Enabling Neighborhood Health Research and Protecting Patient Privacy

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Brittany Marie Krzyzanowski

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Steven M. Manson, Advisor

August 2021

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Steven Manson, for his guidance and support during my scholarship. His endless encouragement, patience, high-standards, and critical feedback helped guide and motivate my research. He provided vital input on various aspects of my work that helped me to gain greater depth in my investigation as well as to better communicate these findings in my writings and graphics. I am also very grateful for his sincere advice, genuine interest in student well-being, and his sarcastic sense of humor—all of which helped me maintain a healthy state-of-mind and stay productive during my studies.

I also thank Michael Oakes who inspired the work in chapters 2 and 3 of this dissertation. His enthusiasm and encouragement gave me the confidence to pursue questions that I might not have otherwise, and his passion for promoting practical, effective, and transparent research endeavors helped to direct the way I carried out this dissertation and will continue to shape my work as I go forward. I am grateful for his eager, honest, and forthright mentorship which energized my curiosity and motivated me to bring this dissertation to fruition.

Special thanks to Mark Lindberg, for his guidance, wisdom, and advice during the duration of my dissertation work. I am grateful to have been able to rely on his cartographic expertise and knowledge of GIS methods and theory, especially during my first few years as a graduate student. His ability to identify critical holes and build thoughtful responses to every concern helped strengthen this project. Additionally, I am thankful that he underscored of the importance of GIS programming and encouraged me to expand my skills in Python and R which enabled me to carry out this dissertation.

Completion of this dissertation, and especially the work within Chapter 4, would not have been accomplished without the help of Eric Shook, who provided me with much guidance, assistance, and advice through his teaching and mentorship. I am grateful to have explored Max P regionalization for the first time as a part of a project within his seminar course. Working with this simple script gave me the confidence to pursue other, more complex coding endeavors that were vital to the completion of this dissertation. Moreover, I am thankful for Eric's kindness and approachability as well as his honesty and encouragement—all of which helped to motivate me to learn and grow in new directions.

My sincerest thanks to Len Kne for connecting me with a broad array of population health GIS projects when I was an RA at U-Spatial. Many of these projects played integral roles in the development of my dissertation, including the Ask About Aspirin project which introduced me to regionalization and the pneumococcal pneumonia mapping project which introduced me to the issues around patient data privacy. I am forever indebted to U-Spatial which is where I garnered the bulk of my research and professional experience. In fact, as of today, nearly a third of my CV entries are affiliated with work at U-Spatial, from publications to grants to presentations to volunteer work. These opportunities were vital in building my skills and facilitating the progress of this research.

I am very grateful to the lawyers, compliance officers, and privacy experts who offered their time to speak to me about their experiences with protected health information. I also thank those who played brief but critical roles in the development of my work on HIPAA law and mapping including, David Van Riper, Amanda Clark, and Zane Wagner, as well as those who provided advice and assistance in the development of my methods including Levi Wolf, Lee Croft, Peter Wringa, and Brian Sweis.

I thank the U of MN Academic Health Center and Fairview Health for providing the patient data used for this analysis. I also thank those from Clinical Information

Services, the Office of Health Sciences Technology Research Development & Support, and the Clinical & Translational Science Institute who helped by providing access to data and critical training and assistance with the data shelter, including Gretchen Sieger, Tim Meyer, Karen Baker-James, and Sonya Grillo.

I greatly appreciate the support of my colleagues either previously or presently in the Human-Environment Geographic Information Science group including Melinda Kernik, Bryan Runck, and Chelsea Cervantes De Blois, as well as past and current members of the MGIS program and U-Spatial including Agata Miszczyk, Taylor Long, Geovanna Hinojosa, Carl Reim, Michael Clementz, Kevin Ehrman-Solberg, Jacob Hartle, and Coleman Shephard.

I would also like to thank Dr. Richard Deyo, my first undergraduate advisor, who gave me the confidence to pursue independent research in a field that I found very much intimidating at the time. My experience working with him in the behavioral neuroscience lab at Winona State University helped direct the future of my scholarship and brought me to where I am today.

Special thanks to my twin sister Constance Krzyzanowski-Dent for providing graphic design input and assistance on several maps and graphics featured within this dissertation.

Chapter 1 Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

For over two decades, public health and medical research has been largely dominated by studies that rely on the use of multilevel statistical models for describing the relationship between neighborhood level characteristics and various health outcomes. Multilevel modeling can be a powerful approach but it can be limited in its ability to capture the complexity of human-environment relationships. Our understanding of health and disease is dependent on the analytical and exploratory methods we use, and therefore many have advocated for neighborhood health research to take a more eclectic approach to data analysis. By encouraging a comprehensive research agenda that integrates a broader realm of new and powerful analytical techniques, we step closer to garnering a more complete knowledge of neighborhood effects on health and well-being.

One way to expand neighborhood health research methods is by integrating mapping and other spatial data exploratory procedures into the research process. Geovisualization and spatial analysis can offer a more complete understanding of complex population health relationships by revealing important, hidden nuances that unfold across space. The relevance and utility of maps for exploring neighborhood health in particular can be important considering the inherently spatial nature of neighborhood health research, and yet spatial analysis and mapping appear to remain underutilized in the literature.

Paper 1: Geovisualization in health research

It is important to use and share maps because spatial data visualization offers to make it easier for readers to comprehend complex, dynamic associations (such as those common within neighborhood health research). The role of geovisualization in comprehension is especially important when considering the recent rise in interdisciplinary efforts among various scholarly and professional institutions. Geovisualization can facilitate effective communication within and between academic domains, and effective communication is vital to supporting successful interdisciplinary research. In other words, visualization provides a common language that can guide these interdisciplinary collaborations.

Furthermore, maps may also inspire new hypotheses by more clearly presenting spatial trends, patterns, and outliers that may have been overlooked. Therefore, in addition to improving comprehension and helping to effectively communicate findings, visualization opens avenues for interdisciplinary research and data exploration by offering a means to uncover previously unseen patterns that may inspire new ideas. For these reasons, it is imperative for researchers to integrate mapping and spatial analytics in investigations of neighborhood health and that they further share these results and tools in their publications.

Chapter 2 of this dissertation examines the extent to which maps and spatial analyses appear (or do not appear) within the literature on neighborhood health. This examination is presented in the form of a literature review and focuses on articles published between the twenty years that span 2000 to 2020, which arguably encapsulates both the rise of interest in neighborhood health research as well as growing interest in and use of GIS for mapping. This chapter offers key insight into temporal trends in the proportion of maps present within the neighborhood health literature according to journal type and changes in the level of sophistication of the maps being published over time. A second, but vital, aim of this chapter is to explore authors' motivation, and barriers, to sharing (or not sharing) maps within their publications. This goal is achieved through the use of survey methods and relies on the corresponding authors of the publications used in our literature review. The last major review of the public health literature's use of spatial methods was over a decade ago (Auchincloss et al, 2012). This chapter revisits this topic with a particular focus on the public health studies that look at differences between neighborhoods. By describing the current state of the literature on neighborhood health and surveying the authors in regards to their use of spatial methods, we uncover the kinds of barriers that stand in the way of expanding the use of spatial methods in public health.

Paper 2: Challenges with patient data privacy law

One obstacle that can stand in the way of researchers trying to share maps and spatial data within their publications is the HIPAA safe harbor privacy provision that protects

patient location data. This provision poses several challenges to researchers wanting to use and share spatial data. First, many researchers find core elements of the provision ambiguous or difficult to understand, which is reflected in disagreement and uncertainty in research and policy circles on how to enact this provision. Second, playing it safe by taking a conservative approach to sharing maps in order to better meet safe harbor provisions — most often by releasing only highly aggregated maps or no maps at all — is a form of data loss that imposes potentially serious costs because it does not allow for the examination of local health distributions at reasonable resolutions for many common health problems. These two challenges have led to disagreements about how to follow the rule and, in fact, the literature is spotted with examples of scholars describing the tenets of the privacy provision in ways that are misleading or using patient location data in ways that are not in compliance with HIPAA law. At the same time, it is not unreasonable to suggest that data could be safely shared in some of the ways described by these scholars. In fact, the literature on data de-identification often challenges the safe harbor provision, saying that it is possible to share finer-grained mapped health data without jeopardizing patient privacy.

One of the major contributors to the barriers observed in Chapter 2 (Paper 1) was the privacy regulation specific to sharing spatial data. For this reason, Chapter 3 (Paper 2) addresses how privacy regulations, specifically the safe harbor rule, hinder the ways in which epidemiologists and geographers understand how to share spatial data. This chapter draws from existing research in data privacy, de-identification, and reverse engineering, as well as congressional records, legal guidance documents, and interviews with compliance officers and lawyers with expertise in HIPAA law to elucidate the ambiguities that burden those trying to understand the privacy provision specific to spatial data. The aim of this chapter is to shed light on how the law was created and how it has been understood (and arguably misunderstood) over that past two decades. This chapter concludes with discussions on how alternative methods to safe harbor can offer researchers better data and better data protection.

Paper 3: Regionalization as a way forward?

One promising way for researchers to share finer-grained mapped health data without jeopardizing patient privacy is with *regionalization*. Regionalization is a geospatial analytical process that builds custom regions from underlying data to suit a specific function or for the display of specific data. This approach gives researchers control over the shape, size, and demographic makeup of the resultant regions within their map. Regionalization holds many potential advantages for the analysis and sharing of PHI under the guidance of HIPAA's safe harbor provision. This is because the population requirements of the safe harbor provision, along with other requirements, can be integrated into the regionalization procedure so to make the maps we share more useful while still maintaining a sufficient level of protection. In spite of the general importance of regionalization to spatial analysis, its use in the context of privacy protection has been very limited. Only a small handful of studies have explored the use of regionalization as a means to create units that meet data privacy regulations and these few studies provide little in terms of publically available tools and workflows that epidemiologists and geographers could easily use. There is a real need for a greater variety of ways to work with, present, and understand patient health data and neighborhood health researchers have much to gain from exploring regionalization as a means to better represent and share protected health information.

Chapter 4 (Paper 3) explores regionalization as one promising way to protect health data while at the same time making it more useful. In this chapter, regionalization, or zone design, is used to build geographical units within the Minneapolis metro area for modeling and displaying data on depression risk. This chapter draws from existing literature on spatial data partitioning, cluster detection, and neighborhood assessment in order to advances knowledge of how regionalization can be used to analyze and report protected health data in ways that satisfy the population threshold delimited by HIPAA guidelines. Four different regionalization approaches are explored for their ability to develop more meaningful units for the display and analysis of patient data. Two of these approaches are novel variants that integrate self-organizing maps into the regionalization process. Our case study uses a real public health dataset (depression diagnoses) to assess

best-fit among different regionalization outputs. Therefore, in addition to advancing the theory and method of data-sharing and visualization, there is potential to provide innovative tools to facilitate dissemination of fine-scale information within patterns of depression to the community.

In sum, the three papers in this dissertation together bring attention to a problem with the way researchers understand how to use and share spatial data and offer a solution in the form of guidance, strategies, and workflows that can help investigators work within the bounds of privacy provisions to share maps and spatial data. Specifically, this dissertation brings attention to the deficiency of maps and spatial analyses published within the literature on neighborhood health and points to the ambiguous rules that guide how researchers can share geographic data as a potential cause for confusion. This dissertation also offers clarity and guidance in the form of a detailed examination of the safe harbor rule specific to geographic data and presents a number of regionalization strategies as a way to work flexibly within the constraints of the safe harbor rule.

Chapter 2. Where are the Maps in Neighborhood Health Research?

Abstract

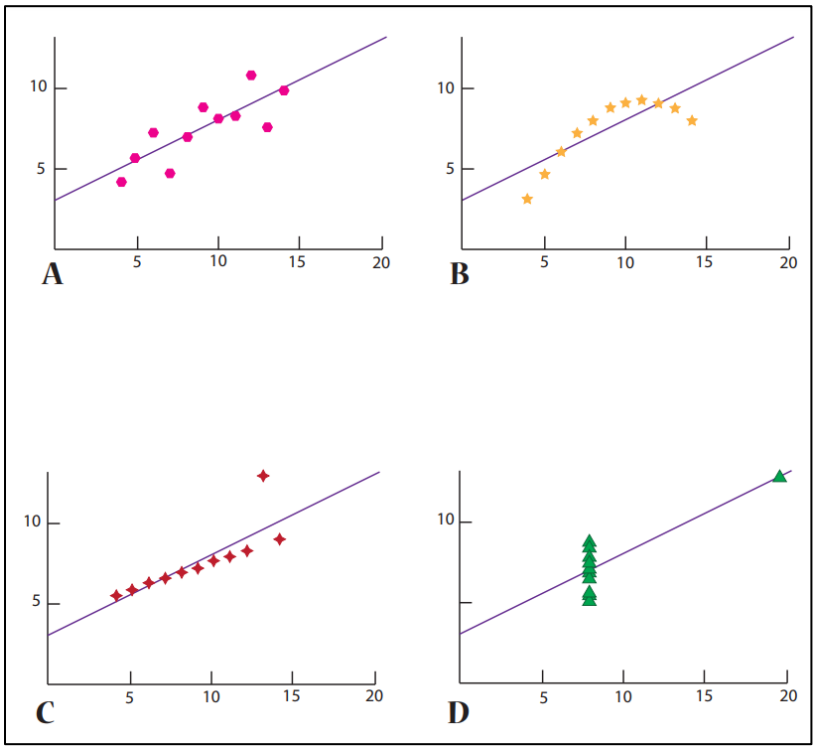
Introduction. Despite large and growing interest in using spatial data and analysis in health research, there appears to be remarkably few maps within the neighborhood health literature. Just as data visualizations, such as scatter plots and histograms, are vital to the initial steps of data analysis, so too are maps for scholarship and policy on neighborhood health. **Methods.** A review of 233 articles on neighborhood health published between 2000 and 2020 was used to identify the proportion of maps within the literature, and a subsequent survey was conducted to identify authors' motivation, and barriers, to sharing (or not sharing) maps. We analyzed temporal trends by journal type and map complexity, alongside the survey results. **Results.** Of the 233 articles reviewed, 64 contained maps. The proportion of maps found within the literature steadily increased over time for both health science and geography/social science journals with the greatest proportion of maps appearing within the last half decade. There were a growing number of higher-level and more sophisticated maps alongside the general increase in maps observed over time. We invited the authors of all papers to complete a survey on map use and sharing and 64 were completed in full. Interestingly, the majority (63%) of investigators created maps or used mapping software to explore questions of neighborhood health but only a small proportion of the maps created by investigators were actually shared within their publications (29%). Survey results indicated that the primary reason for abstaining from sharing maps was the belief that a map would not add value beyond what was provided by statistical models. Other common barriers included journal restrictions, time constraints, and HIPAA or other privacy regulations. The survey indicated that most authors (>80%) reported results in the form of point estimates from regression output. **Conclusion.** While correlation or regression coefficients do a good job at describing the general strength and nature of how two variables coexist in space, maps are needed to understand important, hidden nuances unfolding across neighborhoods. Fortunately, even though maps do not appear frequently within the literature, the majority of studies of neighborhood health use GIS in some way, shape, or form and this figure appears to be increasing over time.

1 Introduction

Data visualization has long been a fundamental step in many forms of exploratory data analysis. This need for, and use of, visualization holds especially true for public health and epidemiology because they rely on, and offer fundamental advances in, applying sophisticated statistical methods. Research in the public health subdomain of neighborhood health, given its focus on the importance of space and place in health dynamics, uses a mix of visualization approaches that includes mapping spatial data. Despite this spatial focus, only a minority of papers using and analyzing spatial data to understand neighborhood health employ maps as a final product in publications. We set out to understand our anecdotal sense of why there appears to be so few maps. Maps offer to enhance neighborhood health research by revealing complex spatial dynamics that lead to new insight and discoveries. They also promise to usher the field towards more ambitious and effective research, one that focuses on using creative and novel approaches to assess the synergistic relationship between neighborhoods and health.

1.1 New insight and discoveries

The importance of “plotting the data" is taught early and often in quantitative methods courses and for good reason. Plotting data is among the easiest ways to uncover the nature of relationships that could otherwise go unnoticed. Consider Anscombe’s Quartet (Figure 2.1) in which the seemingly same linear relationship is expressed by four very different sets of data (Anscombe, 1973). One of the simplest ways to notice the influence of outliers and trends in a set of data is by visually plotting the data, and this simplicity is why exploratory data analysis is an essential first step in the research process. Many public health studies predominantly rely on point estimates, such as regression coefficients, to describe risk relationships, but they also tend to visualize data in the form of scatterplots, box plots, or histograms. For many studies, these forms of visualization are good enough. For public health studies that focus on the spatial relationship between health and neighborhood exposures, however, the data are spatial and a valuable form of visualization is a map.



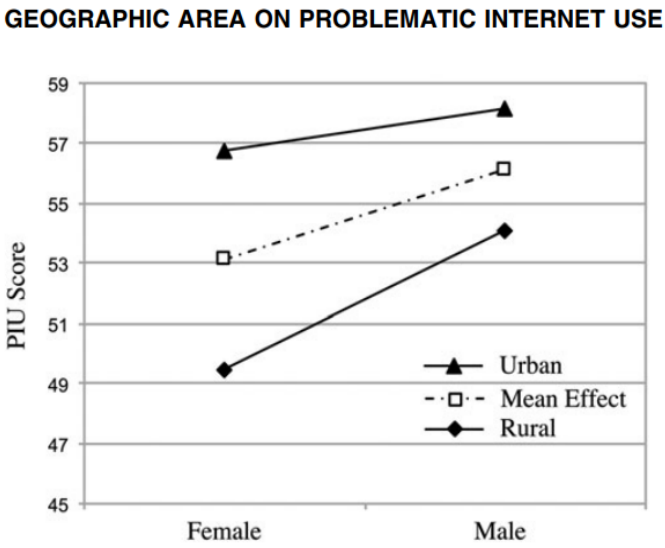
2. 1 Anscombe's Quartet after Anscombe (1973).

Maps are a type of geographic visualization, or geovisualization, which encompasses anything from a simple paper map to more complex, dynamic, interactive web maps or three or four dimensional figures. Maps provide much of the same kind of insight as scatter plots and box plots by allowing researchers to picture the potential influence of outliers, trends, and clustering in their data. Spatial data visualization enables the simultaneous investigation of multiple factors across space and time and therefore maps are key to understanding the drivers of relationships within neighborhood health. Maps have great potential to improve studies of neighborhood health by providing new and powerful ways to analyze and explore data such as with spatial clustering and autocorrelation analysis, assessment of movement and trajectories, agent-based models, and social network and activity space research (Dodge, 2021; Page, 2008; Entwisle, 2007). There are a slew of opportunities for geovisualization to improve neighborhood

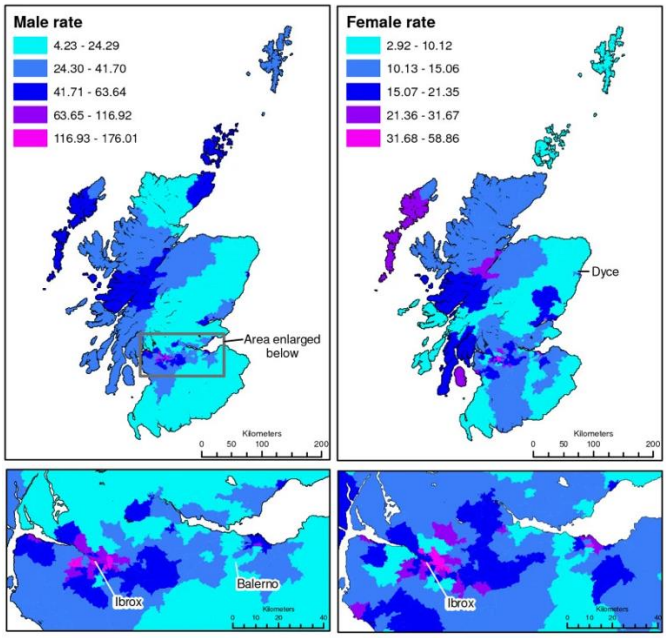
health studies, but one of the most crucial is exploratory spatial data analysis that can lead to new insights and discoveries.

One example can be found in considering Liang Yu and colleagues' investigation of gender differences in adolescent problematic internet use (PIU) which found that men scored higher on PIU than women (Figure 2.2) (2018). If one were to map the gender differences, it would become apparent that this effect is moderated by geographic area (specifically urban and rural). Without considering spatiality, this gender/geography interaction would remain hidden within the mean effect unless investigators had the foresight (or interest) to examine the influence of urban and rural environments on PIU (as these authors did). These scenarios are not uncommon and many have found health and disease to be moderated by geographic area in ways that are often not apparent when just using summary statistics or point estimates (Sheu-jen et al., 2010; Zhang et al., 2014; Gu et al., 2015; Van Os et al., 2001). In cases like these, maps provide deeper insight into the broader picture of the influence of place on health and wellbeing.

Additionally, mapping can provide insight into more complex place-based dynamics such as those presented in Emslie et al's 2009 paper on gender differences in the geography of alcohol-related mortality in Scotland (Figure 2.3). Close inspection of the alcohol-related mortality rate maps reveals marked, place-based gender differences in mortality, yet unlike the Yu et al. work, no clear rural/urban distinction emerges from these maps. There are other, potentially more complex underlying place-based factors moderating gender-based risk for alcohol-related mortality including differences in local labor markets, community culture, and gendered experiences. In addition, maps can be used to identify more than just the obvious geographical splits like urban/rural differences because they can also be useful in identifying less overt place-related influences such as traffic volume (Cakmak et al., 2012), altitude (Beall 1981), or industrial noise (Stansfeld et al., 2000) that might be moderating relationships between the explanatory variables and disease risk.



2. 2 Interaction between gender and geographic area from Liang Yu et al. (2018).



2. 3 Alcohol-related mortality rate for men and women in Scotland from Emslie et al. (2009).

1.2 Ambitious and effective research

In addition to generating new insight, maps may also help to encourage more ambitious and effective research by presenting neighborhoods as something greater than contextual components or activity containers. Maps, especially dynamic, interactive maps, reveal the complex interplay between multiple overlapping risk factors across space and time (Dodge, 2021). This means that neighborhoods are emergent, or in other words, born from the interactions between people and places, and thus geovisualization can prompt investigators to do more than rely on arbitrary boundaries, such as census tracts or county lines, to delineate observational units for a regression model. Acknowledging the complexity of neighborhoods requires that researchers explore a greater variety of analytical methods that offer new and innovative ways to examine the relationship between neighborhoods and health.

For over two decades, public health and medical research has been largely characterized by studies that rely on the use of multilevel statistical models that can be limited in their ability to address confounding in observational data (Oakes et al, 2015; Bingenheimer and Raudenbush, 2004). As a result, many have called for public health researchers, and specifically neighborhood health researchers, to consider taking a more eclectic approach to data analysis (Oakes et al, 2015; Diez Roux et al, 2010; Entwisle, 2007). Geographic visualizations, including maps, are but one among a myriad of available approaches that offer to advance studies of neighborhood health. The relevance and utility of maps for exploring neighborhood health seems obvious given the inherently spatial nature of this kind of research, and yet spatial analysis and GIS remain very much underutilized in the public health domain (Auchincloss et al., 2012; Jacquez, 2000); this is in spite of the fact that neighborhood health research and mapping have complimentary histories.

The rise of neighborhood health research in the 1990s coincided with the introduction of geographic information systems (GIS) that are used for the production of maps and for spatial data analysis (Diez Roux et al, 2010). GIS continued to grow in popularity into the

2000s and manifested as numerous open-sourced mapping platforms (i.e., GeoDa, QGIS, and OpenStreet Maps) that offered more analytical choices during a time when the number of neighborhood health studies being published was increasing exponentially (Diez Roux et al, 2010). Despite the growing availability of software and methods alongside the growth in neighborhood health research, it is unclear whether neighborhood health researchers were taking full advantage of the powerful mapping resources available to them. Around this same time (in the early to mid-2000s) there was rising concerns around the effectiveness of neighborhood-level health interventions and the misestimation of neighborhood effects (Oakes, 2004; Diez Roux, 2004; Subramanian, 2004; Didelez and Mendelian, 2007).

In the fifteen years following the turn of the century, a slew of studies on neighborhood health found that health and disease were spatially organized. This work established that areas characterized by social, environmental, or economic deprivation were generally associated with poorer health outcomes on a number of different measures including overall mortality (Bosma et al., 2001), chronic disease risk (Freedman et al., 2011), infectious disease risk (Iroh Tam, 2017), and mental health (Mitchell et al., 2015). However, a handful of these kinds of studies were being challenged for focusing on the identification of independent neighborhood effects (Oakes et al, 2015), relying too heavily on multilevel statistical modeling (Diez Roux and Mair, 2010), or failing to prioritize consequentialist research questions (Nandi and Harper, 2015). The core of the spatial problem is that multilevel regression models cannot alone piece apart the independent effects of neighborhoods on health in observational studies, and therefore causal claims cannot be made and effective interventions cannot be developed. This shortcoming of neighborhood health research becomes more concerning after noting that targeted community health interventions and preventive efforts driven by neighborhood research are scarce, and the few studies that exist have only shown modest results (Oakes et al, 2015; Diez Roux et al, 2010; Nandi and Harper, 2015).

Instead of casting doubt on neighborhood health research altogether, many have advocated for simply changing the way neighborhood health research is carried out (Chaix, 2009; Nandi and Harper, 2015). Studies of neighborhood health are vital (regardless of independent place effects) because they offer insight into the complexities of communal health by acknowledging the synergy between compositional and contextual neighborhood effects (Oakes et al, 2015; Diez Roux, 2010). As neighborhood health investigators, we need to acknowledge that neighborhoods are inherently spatial, and that exploring neighborhood health data should involve geovisualization of some sort. Put another way, spatial regression, without apriori exploratory spatial data analysis, is the equivalent of a spatial Anscombe's quartet, where important nuances remain hidden within the data if never mapped. For this reason, the generally lackluster success of neighborhood intervention efforts may not actually be so surprising. In order to address these limitations and pursue a more comprehensive research agenda, one must integrate a broader realm of new and powerful analytical techniques that take into account the multidimensional nature of neighborhood data.

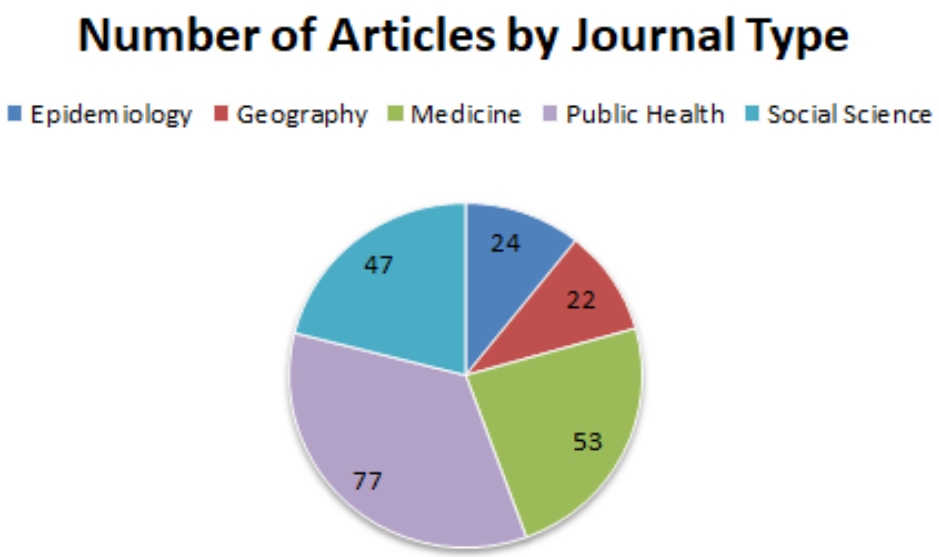
Despite the advantages of mapping and spatial analysis for health research, our anecdotal sense was that there seems to be fewer maps and less spatial research than expected within the literature on neighborhood health, but it is unclear to what extent this sense is accurate. It is clear, however, that our understanding of health and disease is limited by the analytical and exploratory methods we use. Therefore the aims of this paper are twofold: 1) to describe the extent to which maps are present within the literature on neighborhood health, and 2) to assess motivations for, and identify barriers to, sharing maps. This paper sheds light on the current state of the literature and encourages neighborhood health researchers to explore the geographical nature of their inherently spatial data.

2 Methods

In the following paragraphs we describe the strategies taken for carrying out our literature search and subsequent survey. Although not exhaustive, our review of the literature was substantive enough to gather a good understanding of the current state of the literature, having searched 600 articles and identifying 233 that could be included in our analysis. For these articles, we designed a brief survey to assess author's motivations for, and barriers to, sharing maps.

2.1 Article Selection

We selected 233 articles (table in appendix) from 103 different journals (Figure 2.4) via an electronic keyword literature search on Google Scholar. The articles were selected from the first hundred results in six separate Google Scholar searches (600 search results in total) of articles published between 2000 and 2020 with the keywords "neighborhood health" and either "census tract", "block group", "ZIP code", "county", "city" or "municipality". Neighborhood health captured a very broad range of topics in this review; we considered articles ranging in topic from infectious and non-communicable disease to sleep disorders to fire injuries to food security to criminal behavior. Articles were only included in the analysis if they contained a map or had the potential to contain a map (we refer to this as "map potential" for the duration of the paper). For example, an article had map potential if it used a dataset that contained information on individual or aggregate location, or performed a door to door or telephone (landline-based) survey. Additionally, we were also careful to exclude articles where a map would not contribute to the aim of the study (i.e., it makes sense to map patient addresses, but it wouldn't make sense to create a map of simulated data used within a methods paper). One of the ways we ensured map potential was by requiring studies to span a large enough area to be considered more than one "neighborhood" (in our case, neighborhoods ranged from a small number of blocks to a large number of entire cities). Perceptual neighborhoods were included in the study if individuals who provided their perceptions were also linked to an aggregated level of geography that could be easily used within an analysis.



2. 4 The number of articles from each journal category. Note that ten articles were classified as "other" and were excluded from journal category assessments.

We acknowledge the extent to which our methods can be considered subjective in the sense that other investigators may have selected different keywords or explored deeper into the search results. Our strategy was to try to capture a broad array of relevant papers without claiming to be exhaustive. The selection criterion was specific enough to leave little question in relation to which articles should and should not be included within the study, and the sample of articles was broadly representative of the neighborhood literature at this point in time. This being said, we discuss strategies for future work in the conclusion. In particular, it is important to note that Google Scholar provides search results that are meant to replicate how researchers rank results, which means it is weighted articles according to how well the article is cited in other scholarly literature. Following the advice of Haddaway and others (2015), the current project is not meant as an exhaustive and systematic review but instead an initial investigation of articles using one of the primary tools that many researchers use in exploring academic literature. The authors acknowledge the extent to which some studies have found traditional academic searches, such as those using the Web of Science platform and PubMed, to provide better precision and to be less biased against grey literature compared to Google Scholar when reviewing science and biomedical topics (Haddaway et al 2015; Anders et al., 2010).

However, when the topic to be reviewed was a social science topic (elderly migration), one study found Google Scholar to provide better general performance (in terms of precision and recall) compared to MEDLINE, Academic Search Elite, Social Sciences Abstracts, and EconLit (Walters et al, 2009). For this reason we find Google Scholar to be a suitable platform for our review of the literature on neighborhood health which often times centers on the intersection of public health and social science.

2.2 Map Definition

Maps were defined as a figure or graphic that contained locational information in such a way as to illustrate the distribution of a health outcome or risk factor across two or more geographical areas. We expected maps to fall within one of three commonly-accepted geometric categories: 1) aggregate or polygon maps (e.g., census tracts, counties), 2) point maps (e.g., locations of liquor stores, health clinics), and 3) line or network maps (e.g., road or social networks). No line or network maps appeared in our study sample.

2.3 Survey Format

After collecting articles from the electronic keyword literature search, the corresponding authors of the articles were contacted via email and asked to complete a short survey. The survey was created and administered anonymously within Qualtrics and therefore information was not collected if it could be used to link respondents to an article on our list (such as journal name and publication date). The survey was four to six questions and took less than one minute to complete. The survey included multiple choice and open-ended questions and only allowed one response per question. The IRB determined that these activities were not research involving human subjects as defined by DHHS and FDA regulations. Participation was voluntary and therefore our analyses and discussion of survey results only consider the responses from individuals who provided consent (Figure 2.5).

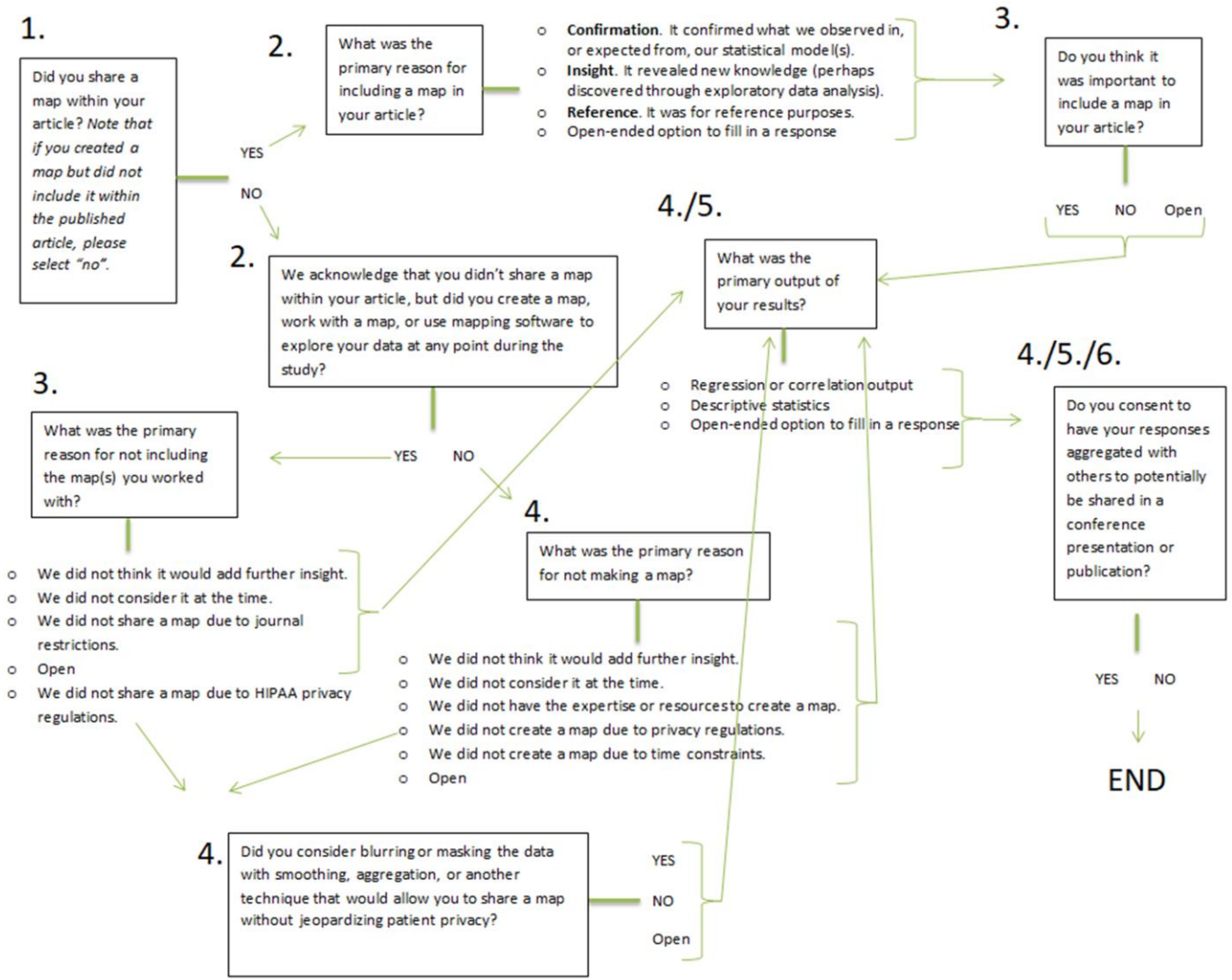


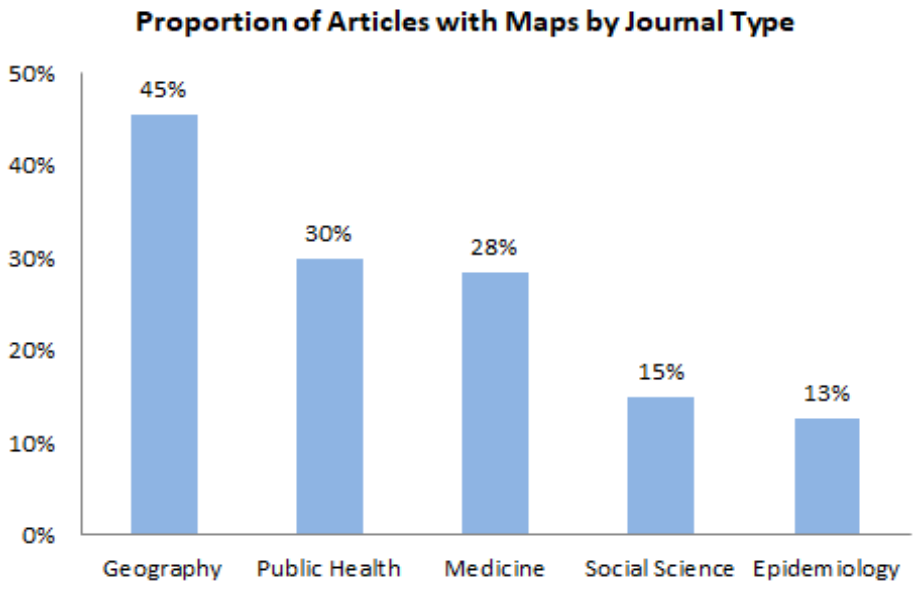
Figure 2.5 The full survey

3 Results

In the following paragraphs we present the results from our literature review and subsequent survey. Results from our literature review include an examination of the proportion of maps and spatial analyses found, as well as an evaluation of the differences observed across time after stratifying by journal type and map complexity. Our survey had a response rate of 31% (which is impressive for external surveys) and provided a sufficient sample from which we could gather insight. All results are reported in the form of proportions, charts, and graphics.

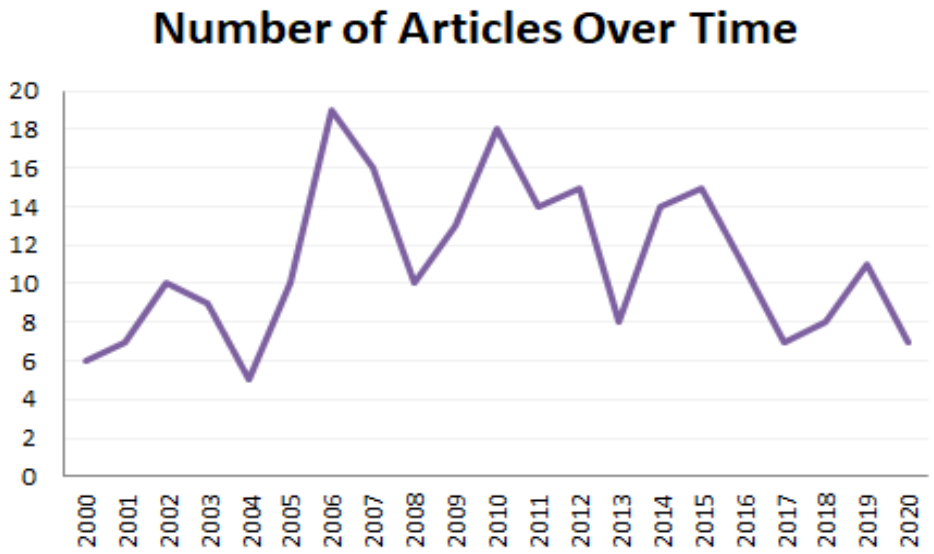
3.1 Literature map analysis

Our sample contained 233 articles on neighborhood health published between 2000 and 2020. Of these, 64 articles contained maps (27%), which is to say that the majority of articles on neighborhood health (73%) did not include a map. Furthermore, the presence of maps varied with journal type. Of the sample of articles collected, 154 out of 233 were pulled from public health, epidemiology, and medical journals while the remaining articles were classified into social science, geography, and general science journals. In terms of maps, 30% of articles from public health journals, 28% of articles from medical journals, and 13% of articles from epidemiology journals included at least one map (Figure 2.6). Geography journals had the highest proportion (45%) of articles that included maps.

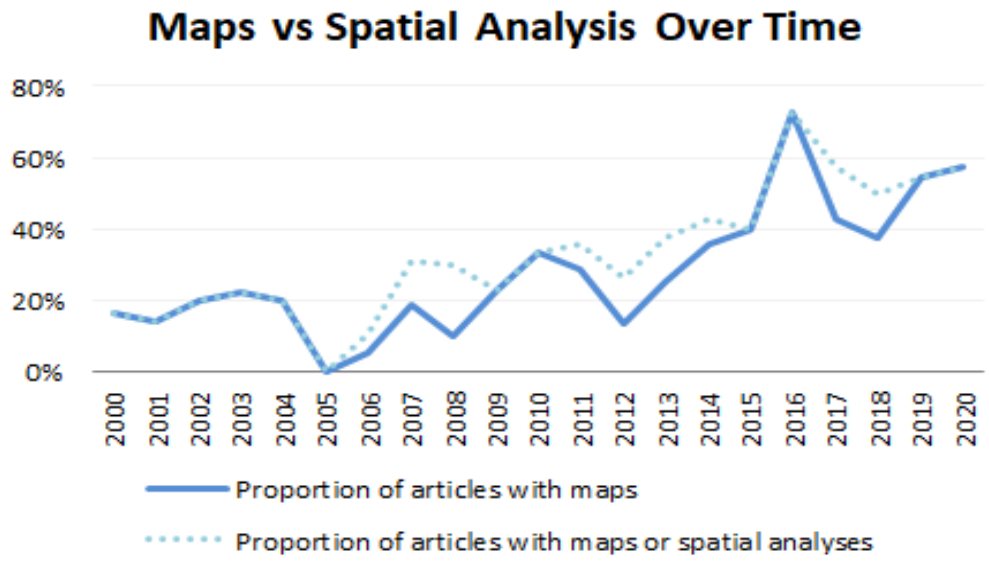


2. 6 The proportion of maps published within the neighborhood health literature according to journal type.

In addition to exploring the prevalence of maps in the literature by journal type, we also explored trends over time. Based on our sample, the number of neighborhood health publications peaked around 2006, with the number of publications nearly tripling that year (Figure 2.7); the years since then have seen fairly consistent production of papers. As noted above, Google Scholar results are sorted by relevance and not by date, which means that more recent papers may be subtly discounted. The proportion of maps present within the literature on neighborhood health seems to follow a general trend of increasing over time (Figure 2.8). To supplement this finding, we identified the articles that did not contain a map but did perform some sort of spatial analysis (i.e., spatial lag models, spatial CAR and SAR, spatial autocorrelation). Doing so allows us to better understand the extent of the investigator's awareness of spatial analytical techniques. However, of the papers that did not contain maps, we found that only a small proportion (7%) performed some sort of spatial analysis. Still, when these data are plotted (as the proportion of articles containing maps or spatial analyses) over time we reveal a much more stable upward trend (Figure 2.8).



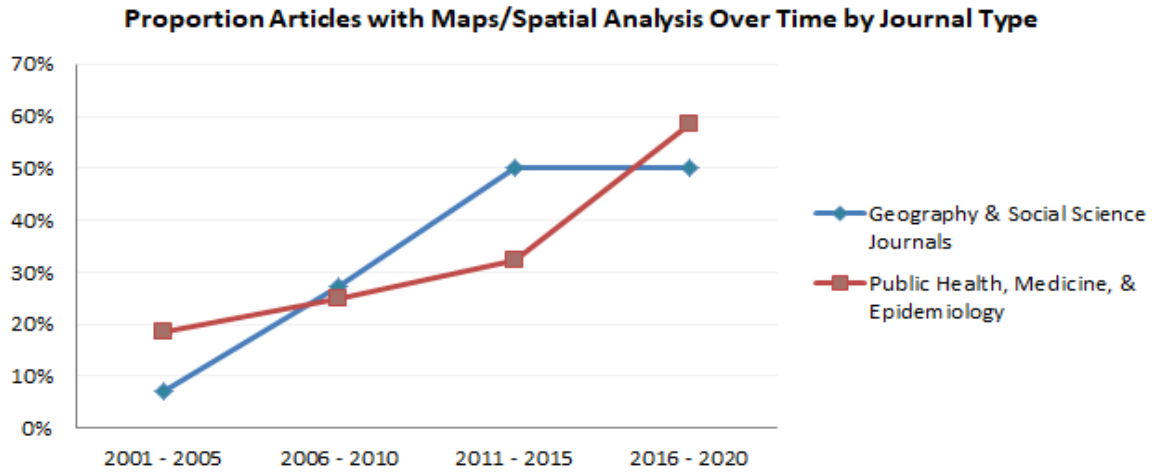
2. 7 The number of neighborhood health papers published by year (based on our sample of 234 articles).



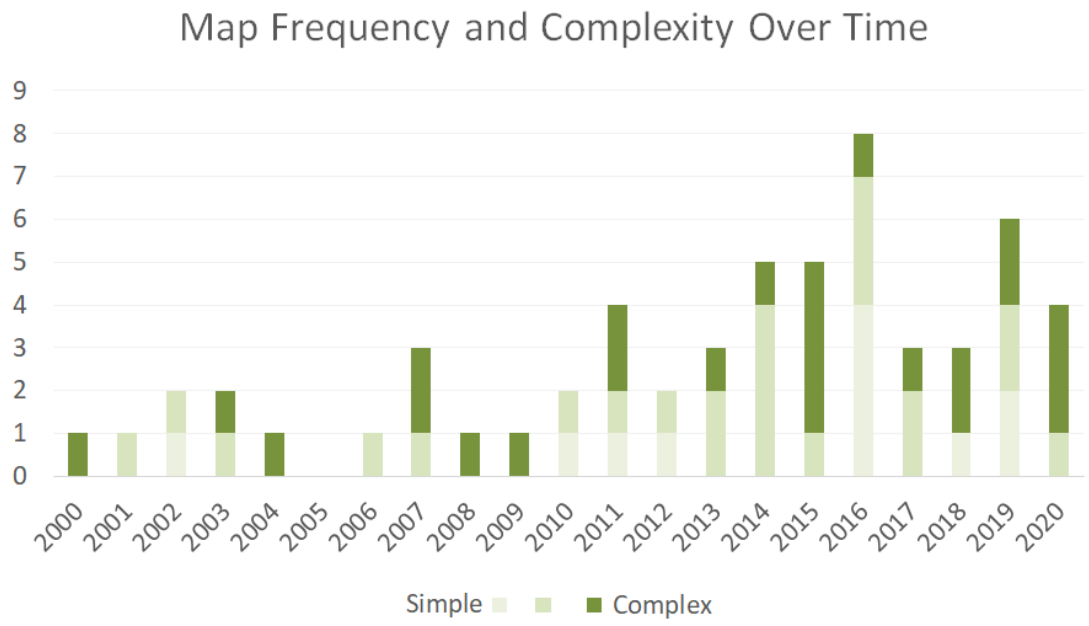
2. 8 The proportion of articles from our sample containing maps and the proportion of articles containing maps or spatial analyses published between 2000 and 2020.

In order to explore temporal trends by journal type, our data was first aggregated as to provide more reliable rates due to some journal categories (e.g., geography and epidemiology) not being as well-represented as others (e.g., public health and medicine) within our literature search (Figure 2.9). Accordingly, time is represented in 5-year increments and the journal categories were merged to form two main categories: 1) Geography and Social Science (GSS) and 2) Public Health, Medicine, and Epidemiology (PHME). When comparing temporal trends between the two main journal categories, we found both categories to exhibit a general increase in the proportion of maps published over time. However, for the PHME journals, this increase was relatively gradual until a spike in the last half decade, whereas for GSS journals the trend was steep and consistent from 2000 through 2015 but it dropped off thereafter. These trends become more apparent when comparing the proportion of articles containing either a map *or* spatial analysis (Figure 2.9), which illustrates how in the last half decade the PHME journals exhibit a spike in the proportion of maps or spatial analyses while the GSS journals flatten out.

Additionally, we explored these trends according to the level of sophistication of the maps being published over time (Figure 2.10). We rated maps as being simple (1), medium-complex (2), and complex (3) according to some simple rules. Simple maps included reference maps and maps of study sites. Medium-complex maps included some form of analysis, usually using choropleth mapping methods. Complex maps included cluster maps (i.e., LISA), time/distance maps, and choropleth maps overlaid by clustered features. Review of the maps published revealed a growing number of higher-level, more sophisticated, maps appearing within the literature, but this is likely a function of the general increase in maps observed over time.



2. 9 The proportion of articles that published maps or included spatial analyses within Geography & Social Science journals compared with Public Health, Medicine, and Epidemiology Journals from 2000 to 2020.

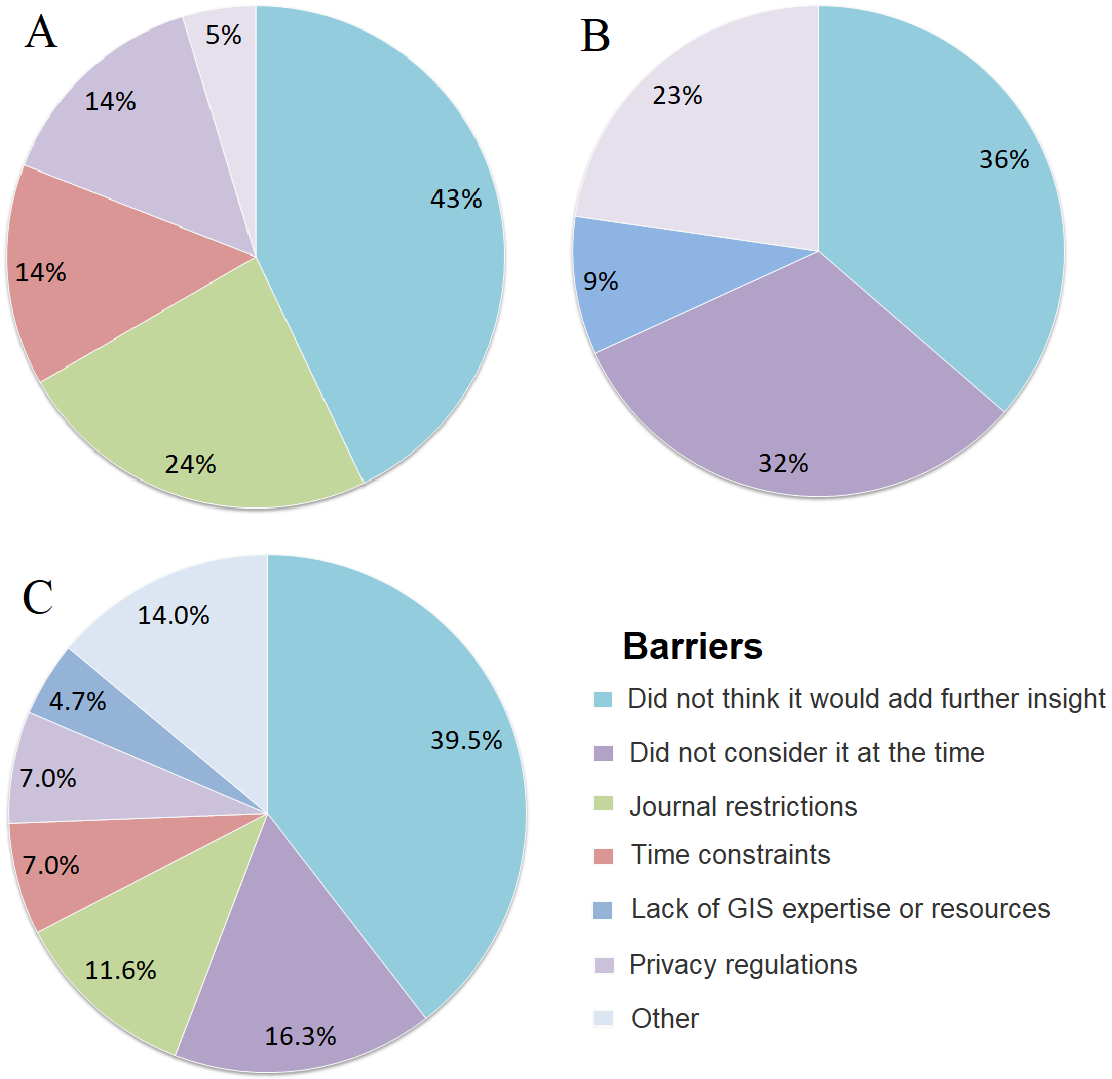


2. 10 Map frequency and complexity across time.

3.2 Survey results

We successfully distributed 207 surveys to corresponding authors from our list of 233 articles. For the remaining 26 authors, our email failed to find a recipient and no alternative email address could be found. Of the surveys sent, 66 were returned, and 64 were completed in full and provided consent to be used within the following analysis. Of the 64 survey respondents, 70% (45) did not share a map in their publication. However, a notable proportion (nearly half) of those who did not share a map indicated that they created a map, worked with a map, or used mapping software to explore their data at some point during their study.

When asked for the primary reason for not sharing the maps that they worked with during their investigations, most of these respondents (43%) thought that a map would not add further insight to their study (Figure 2.11A). The second most frequent response (24% of respondents who worked with but did not share maps) selected journal restrictions as the primary barrier (i.e., paying extra to include a color figure). Other barriers identified range from time constraints (14%) to privacy regulations such as HIPAA law (14%).



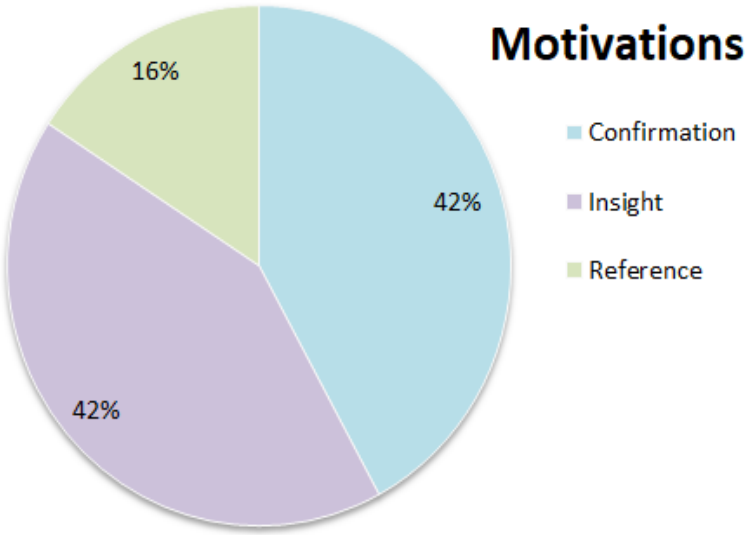
2. 11 A) The primary reasons for not sharing the map(s) worked with during the investigation. B) The primary reasons for not making a map at all. C) The primary reasons for not making a map or for not publishing the map(s) worked with during the investigation.

Of the 22 respondents who did not use mapping software at all, the majority (38%) indicated that they did not make a map because they did not think it would provide extra insight beyond that which was provided by their statistical models (Figure 2.11B). The next most common response (33% of these respondents) selected that they had just neglected to consider mapping their data at the time. Less than 10% of these respondents suggested that the primary reason for not using mapping software was due to not having

the resources or skillset to create a map adequate enough for publication. There were a number of respondents who selected “other” as their primary barrier and many of these respondents did not choose to fill in a more detailed response. Those who did offer more detail wrote that author preferences played a role or that they included maps and/or spatial analyses in other, follow-up publications.

Figure 2.11C shows all of the barriers aggregated for those who did not create maps and those who created but did not share maps. More than half of these respondents did not see the value in including a map (39.5%) or neglected to consider it at the time (16%). Only 5% of respondents considered lack of resources or mapping expertise as a barrier. A sizable proportion of respondents selected the “other” answer choice (14%) which indicates that our survey was limited in its ability to capture every barrier to sharing maps within publications of neighborhood health.

The majority of those who shared maps indicated that the primary purpose for including a map was for confirmation (42%) or insight (42%), being that the map(s) revealed new knowledge or supplemented and confirmed what was observed within the statistical models. Only 3 of the 19 survey respondents who published a map did so only for reference purposes. Survey respondents did not indicate any other motivations beyond these three options.



2. 12 The primary reasons for including a map within the publication.

4 Discussion

It is good practice to explore data before running it through a model. For this reason, we expected that exploring neighborhood health could likely involve creating a map at some point during the investigation process. And, in fact, it seems that this is most often the case. According to the survey, the use of maps in investigations of neighborhood health is relatively common, in that survey results showed that the majority (63%) of investigators created maps or used mapping software to explore their data. Interestingly, the presence of maps within the literature is much lower—only 27% of the 233 articles reviewed in this study included a map and a similarly small proportion (29%) of survey respondents shared their maps. Clearly, public health investigators are not neglecting to explore the spatial nature of their data, but rather, they are just not publishing the maps that they are using.

The reasons for not sharing maps that are being created during investigation ranged broadly, but it is significant that the majority of these survey respondents reported that they thought a map would not add further insight to their study beyond that of which was provided by their statistical models. Related, the majority of people who *did* share maps indicated that the maps were shared for confirmatory purposes, namely to help convey what was observed from the statistical models. Although the majority of respondents indicated that their papers could have still been published without their map(s), nearly all of the respondents reported that they believed it was important for them to share their maps. That being said, an equally large proportion of respondents indicated that they shared a map because it revealed new knowledge likely gained from exploratory data analysis. In these instances, maps demonstrated patterns and relationships that were beyond what could be gathered from their statistical models.

A significant portion of neighborhood health studies simply did not consider mapping their data (as indicated on the survey). This finding confirmed our anecdotal sense but it is somewhat surprising given that neighborhood health research is inherently spatial and that GIS had been introduced and was available many years prior to the time period defined for our article search (2000-2020). One potential reason for this finding is that most of our survey respondents came from medicine, public health, and epidemiology backgrounds with a very small proportion of answers coming from geographers. It is important to note that only 22 of our respondents opted to answer the (optional) research background question. Nevertheless, the difference between groups was stark (17 from medicine, public health, and epidemiology; 4 sociologists; 1 geographer), and this may explain why a notable proportion did not consider mapping their data.

Only a small proportion of our survey respondents identified lack of resources and GIS expertise as a barrier to sharing maps in their neighborhood health publications. Spatial research hubs are becoming more and more common within research universities and therefore it is easier for health research groups to gain access to free spatial data

consultation on campus. However, the extent to which these spatial research hubs are utilized by neighborhood health researchers is unclear. More research is warranted in order to better understand how this barrier (lack of GIS expertise) has changed or remained stable since 2000.

Perhaps more importantly, it is also the case that a lesser, but nevertheless notable, portion of neighborhood health studies are abstaining from sharing map visuals to avoid dealing with journal restrictions and working with HIPAA or other privacy constraints. By only publishing the point estimates from regression and correlation analyses for example, investigators avoid the burden of adding complex graphics to their publication while also guaranteeing HIPAA compliance when sharing their study results. However, these point estimates are limited in what they can do, and despite criticism from many that warn against relying too heavily on the use of regression approaches for the study of neighborhoods (Diez Roux et al., 2010; Oakes et al, 2015; Entwisle, 2007) the majority of our survey respondents indicated that the primary output of their results came from regression modeling.

Regression and multilevel modeling on their own are valuable for understanding many features of human health but not sufficient for capturing complex, dynamic, interdependent relationships between people and place that characterize neighborhood health (Diez Roux et al., 2010). For this reason, maps and geographic visualizations (especially for studies of neighborhood health) should be used to supplement standard regression modeling to help provide greater insight into the role of spatial clustering, outliers, and trends. “Visualization empowers data science” (Dodge, 2021). And, in an age of big data that offers widespread and growing availability of troves of electronic medical records, much of public health and epidemiological research stands to *become* data science. These health data sciences rely on data visualization for facilitating the interpretation of results, revealing unknowns, and communicating concepts to be shared to others.

The good news is that there seems to be a trend whereby the proportion of neighborhood health articles that are publishing maps is increasing over time. Looking at raw numbers we notice the greatest number of maps during the peak of interest in neighborhood health research (between 2006 and 2010), but proportionally to the rate of publication, more maps were shared in the last half decade. This trend can be best seen by looking at the PHME category plotted in 5-year increments in Figure 2.9. This figure depicts a gradual, but notable, increase in the proportion of maps beginning in 2000. Furthermore, while the number of articles in the GSS category is too small to make a complete assessment, it seems that the proportion of maps being published within these journals is also increasing steadily over time. Additionally, in terms of map complexity, a growing number of higher-level, more sophisticated, maps appeared within the literature in the past half-decade, but this is likely a function of the general increase in maps observed over time. Our study was limited in its ability to explore map complexity (only 64 of our articles contained maps). Future research should focus on collecting a larger sample and parsing this trend by domain to assess the changes in map complexity over time among the health and spatial sciences.

Our survey did not ask respondents to provide the date of publication (in order to ensure anonymity), and therefore we are unable to talk about how these barriers may have changed over time. However, with the data that we have, it is interesting to consider how an author's decision to create and share a map seems to be more strongly tied to whether they see value in geovisualization rather than to lack of mapping skills or lack of access to mapping resources. A follow-up survey that gathers publication date information would be required to assess whether inaccessibility to GIS resources were a more formidable barrier in the early 2000s than in the past half-decade. Additionally, it would be interesting to see how the valuation of map visuals may have changed over time. Even though our survey was limited in its ability to capture changes in barriers overtime, we do believe that the insights gathered here help to shed light on some of the attitudes investigator’s hold in regards to the usefulness of maps in neighborhood health literature.

Further research is needed to gather a more comprehensive picture of investigator's perceptions.

Here we pause to consider, is it really that necessary to share the map visuals used during spatial data exploration? As long as investigators examine the spatial nature of their data with geovisualization and/or geostatistical analyses (which according to our survey, most are) then perhaps it is not that important for the actual map visuals to be published alongside the statistics. In fact, not publishing maps would make the research process easier by cutting out the difficulties of navigating HIPAA privacy law and dealing with journal graphic requirements and restrictions. Be that as it may, let's consider what is kept from the readers when only the statistical output is shared. In other words, what do visualizations provide readers that statistics cannot?

In addition to aiding investigators in interpreting their own data via exploratory spatial data analysis, maps and other geovisualizations help readers to better understand the work being presented. Visuals offer to make it easier for readers to comprehend complex, dynamic associations (such as those common within neighborhood health research) by better conveying the strength and nature of the relationships at hand. This is especially important when considering the recent rise in interdisciplinary efforts among various scholarly and professional institutions (Van der Aalst, 2016). Effective communication within and between academic domains is vital to supporting successful interdisciplinary research. According to Somayeh Dodge, geovisualization and movement expert at University California Santa Barbara, "[v]isualization provides a common language for communication in interdisciplinary research and facilitates the collaboration between domain experts, data owners, and developers of methods" (Dodge, 2021, pg 106).

Furthermore, maps may also inspire new hypotheses by more clearly presenting spatial trends, patterns, and outliers that may have been overlooked by the authors of the

publication. Therefore visualization offers a means to uncover “unknown unknowns” (things we don’t know we don’t know) (Van der Aalst, 2016), especially when the visualizations are shared in a dynamic and/or interactive form (i.e., web maps). This is because visualization exploits the human cognition capabilities and, in doing so, previously unseen patterns can emerge. “Insight is the traditional aim of visualization.” (Van Wijk, 2005). Sharing maps could help inspire new research ideas and open unseen avenues for interdisciplinary research and data exploration.

There are clear advantages to sharing maps—whether they be simple printed maps or complex interactive web maps. What remains unclear is whether these gains are substantial enough to justify the time it takes for researchers to create a map, mask private patient data, format the graphic to journal standards, and potentially pay extra for color printing or a web domain host. Future research into the use of and attitudes towards maps and other geovisualizations in neighborhood health is warranted in order to better guide the field towards one characterized by multimethodology.

There are several ways in which future research can build upon this project. The first being to pursue a larger, exhaustive, systematic review of the literature on neighborhood health since the present study was purely a preliminary investigation that relied on a single search platform (Google Scholar). Although Google Scholar has been shown to provide good performance when the topic to be reviewed was a social science topic (Walters et al., 2009), its performance was less impressive for other topics (Anders and Evans, 2010; Haddaway et al., 2015). Furthermore, Google Scholar was found to be biased against grey literature (articles not published by commercial academic publishers) whereby peak grey literature was achieved after page 80 (Haddaway et al., 2015) (well beyond the limits of our search). For this reason, further review should consider using traditional search methods consisting of multiple platforms and setting deliberate strategies to address biases against the grey literature which may or may not contain more maps than the academic literature. Furthermore, we recommend that follow-up

investigations seek to reach a larger number of corresponding authors with surveys. The present study included only 64 survey respondents in its analysis of which only 18 respondents shared maps. A larger sample is needed in order to gain a better understanding the motivations of those who shared maps. Additionally, we would recommend that a more thorough survey be conducted that includes the collection of information on journal type and publication year. Collection of these kinds of information will require more in terms of privacy safeguards, but they would allow for a richer examination that can address how barriers have changed over time for different domain categories.

5 Conclusion

Many have called on investigators to expand their vision of population health research methods and have further encouraged researchers to explore novel approaches and to use a combination of strategies when investigating neighborhood health (Oakes et al, 2015; Diez Roux et al., 2010; Page, 2008; Entwisle, 2007; Oakes, 2004). One way to expand population health research methods is by integrating spatial data exploration into the research process. Geovisualization offers a more comprehensive understanding of complex neighborhood health relationships and therefore it is encouraging to find that, since the year 2000 more and more, neighborhood health investigations are choosing to explore the spatial nature of their data. Despite this, very few studies actually share the maps they make during their exploratory spatial data analysis. Of our sample of 233 neighborhood health papers published in the last 20 years, only a handful shared maps. The impact of the dearth of maps on neighborhood health research remains unclear.

Chapter 3. Twenty Years of the HIPAA Safe Harbor Provision: Unsolved Challenges and Ways Forward.

Abstract

The Health Insurance Portability and Accountability Act (HIPPA) was an important milestone in protecting the privacy of individuals but its provisions are so vague as to hinder how epidemiologists and geographers share spatial data. In particular, the HIPAA safe harbor provisions are ambiguous when it comes to the use and sharing of spatial data, and the effect of this ambiguity is apparent across the literature on spatial health and has resulted in many entities sharing data at what could perhaps be an overly conservative level while others potentially put patient data at risk. This paper promotes understanding of the HIPAA safe harbor provision by providing a comprehensive overview of the law while also presenting various expert perspectives and relevant studies that, taken together, show how alternative methods to safe harbor can offer researchers better data and better data protection. Much has changed in the twenty years since the introduction of the safe harbor provision, and yet it continues to be the primary source of guidance (and frustration) for researchers trying to share maps, leaving many waiting for these rules to be revised in accordance with the times.

1 Introduction

When addressing many kinds of research problems, maps should generally be shared at a resolution that best portrays the reality of the underlying data. In terms of health and disease mapping, this realism often means wanting a fine-detailed visualization that helps make community-level public health interventions more effective. Geotechnologies offer innovative ways to create these fine-detailed maps and to customize them for the analysis and display of health data. At the same time, however, these data and tools can be dangerous when working with sensitive data, such as patient health records. In particular, scholars must be careful not to share maps with so much detail that individual people can be identified. To prevent identification of patient records, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides guidance on ways to de-identify protected health information (PHI) before it is shared, but HIPAA guidelines are difficult to apply to spatial data.

HIPAA law poses several challenges to researchers wanting to use and share spatial data. First, many researchers find core elements of the *safe harbor provisions* of HIPAA (a set of conditions that define how data can be shared) ambiguous or difficult to understand, which is reflected in disagreement and uncertainty in research and policy circles on how to meet the safe harbor standards. Second, playing it safe by taking a conservative approach to sharing maps in order to better meet the safe harbor standard — most often by releasing only highly aggregated maps or no maps at all — is a form of data loss that imposes potentially serious costs because it does not allow for the examination of local health distributions at reasonable resolutions for many common health problems. These two challenges lead to disagreement about how to follow privacy rules and, in fact, many scholars and policy makers have challenged these rules, saying that it is possible to share finer-grained mapped health data without jeopardizing patient privacy.

Addressing the twin challenges of the safe harbor provisions (ambiguity and data loss) requires an exploration of past and current understanding of how the provisions are

enacted and identifying specific ways in which finer-scaled data may be legally and technically possible. Section 2 of this paper begins this exploration by examining the legal dimensions of HIPAA law from its creation through to current practice. This section looks at the events and concerns that fueled the motivations of those who helped write the safe harbor provisions, with a particular focus on answering the question of why ZIP codes and a population threshold of 20,000 were chosen as anchors for the safe harbor. Section 3 explores the first of the twin challenges, uncertainty, and establishes how some unintentional ambiguity in the law has led to different interpretations of HIPAA privacy provisions specific to geographic data in the public health literature. We focus in particular on how this ambiguity has led to two common but different interpretations across a range of scholarship based on 3-digit and 5-digit ZIP codes, and what this means for mapped data. Section 4 presents and explores data loss, the second of the twin challenges of the safe harbor provisions. The section builds on the previous ones to explore whether there is a middle ground to be found between sufficiency and stringency, asking in essence if there are ways to minimize risk under HIPAA while allowing for more useful maps. Section 5 concludes by presenting the new approaches to de-identification of patient data and discusses ways forward.

This paper advances our understanding, and potential use, of the safe harbor provision of HIPAA law as applied to spatial presented as maps. It is the first comprehensive overview of the long-standing and important conversations around this general topic. By untangling the law and reviewing its history and use, this paper offers avenues to finding safe and more useful ways to share mapped patient data. It also seeks to spur a broader conversation about ways forward that necessarily expand and improve shared understanding of the privacy regulations to encourage researchers to investigate alternative strategies.

2 HIPAA Privacy Act: Zip codes and the 20,000 population threshold

In order to better understand the safe-harbor provision and what it asks of researchers it is best to first understand its origin. Looking at HIPAA in terms of its history and evolution sheds light on how to approach sharing geographic information under the safe harbor standard. We ask two related questions: 1) why do ZIP codes hold such sway over defining the safe harbor rule? and 2) why is the threshold of 20,000 people used to define privacy? Answering these questions clarifies some of the key ambiguities in HIPAA safe harbor and gives insight into why there is so much seeming disagreement within and across research domains. The following section provides a brief overview of HIPAA privacy law before diving into the history of the safe harbor provision to provide insight into the two key ambiguities (the use of ZIP codes and the population threshold).

2.1 The safe-harbor provision

In order to protect patients' privacy, HIPAA limits the ways in which patient data can be shared. Patient data is considered Protected Health Information (PHI) that needs to be kept secure because it includes private medical information along with identifying information such as names, birthdates, addresses, and social security numbers. Address data, in particular, is considered extremely sensitive as it (along with other location data such as longitude and latitude) may be used to pin-point the home residence of an individual. This degree of locational specificity substantially increases the likelihood of identification, if not fully guarantees identification in the case of single-occupant residences. For this reason, patient locations need to be masked in accordance with HIPAA privacy law.

Two standards are specified under the HIPAA rule for de-identifying patient data — the safe harbor standard and expert determination — but former is the de facto standard (Office for Civil Rights, 2012). Expert determination—also termed the statistical standard—is the process by which an investigator masks their data and has a third party

expert determine whether the location masking strategy applied provides a low probability of identification. Expert determination is not frequently used in large part because it is ambiguous and requires unspecified documentation, in addition to placing a good deal of pressure on the third-party expert who is charged with certifying HIPAA compliance. This leaves the safe harbor standard as the most commonly relied upon practice for de-identifying patient data. Its immediate appeal, and primary reason for broader acceptance than expert determination, is that it offers ostensibly clear guidance. The safe harbor standard is the focus of the remainder of this paper.

In essence, the safe harbor method protects patient data by simply removing 18 types of identifiers (Table 1). Many of these elements are straightforward to comprehend and implement, such as not including names, birthdates, and social security numbers. Some of the other elements pose their own challenges in an age of surveillance, such as biometric markers including vehicle license plates and facial imagery. Our focus, however, is section (2) of safe harbor, relating to the patient's location, which is especially relevant to mapping and not surprisingly, the primary source of confusion in applying the safe harbor rule to mapping. The location provision of the safe harbor rule requires a minimum population of at least 20,000 people to be contained within each aggregated geographical unit, and the rule further requires that the only permissible geography (smaller than the state) is a form of ZIP code.

Ambiguity arises when the type of ZIP code isn't specified. Although it seems fairly clear from the text below that the rule intends for investigators to rely on the use of 3-digit zip codes (as compared to 5-digit ZIP codes), not all who read this stipulation see it that way. There are many reasons for this including various misleading representations of the rule found in legal online documentation as well as in literature on public health and disease mapping. The following section explores how ZIP codes have come to play a key role in the safe harbor rule.

Table 1. The key elements of the safe harbor provision

The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed: (1) Names (2) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and the initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000 (3) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older (4) Telephone numbers (5) Vehicle identifiers and serial numbers, including license plate numbers (6) Fax numbers (7) Device identifiers and serial numbers (8) Email addresses (9) Web Universal Resource Locators (URLs) (10) Social security numbers (11) Internet Protocol (IP) addresses (12) Medical record numbers (13) Biometric identifiers, including finger and voice prints (14) Health plan beneficiary numbers (15) Full-face photographs and any comparable images (16) Account numbers (17) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section “Re-identification”]; and (18) Certificate/license numbers

2.2 Why ZIP codes?

If we were to remove ZIP codes from the safe harbor provision there would be no ambiguity in terms of its interpretation because the rule would simply focus on the threshold of 20,000 people to define whether some arbitrary geographical unit is sufficient. So why are ZIP codes written into the law? To answer this, we need to start at the very beginning in terms of how it came into being and understand how the political, social, and technological milieu of the time shaped some core principles and guidelines. ZIP codes were originally not included in the rule but this quickly changed as a result of a mix of happenstance and deliberation. The following paragraphs provide insight into the series of events that led up to the HIPAA safe harbor provision that we understand today, beginning at the proposed bill.

Before HIPAA was law, it was a bill, specifically bill H.R. 3103 of the 104th Congress from 1995-1996 (H.R. 3103, 1996). This bill was introduced in the spring of 1996 as part of an initial attempt at healthcare reform by the Clinton administration. The overarching focus of H.R. 3103 was to improve access to healthcare and address fraud, waste, and abuse in health insurance and healthcare delivery, but it also—quite briefly—mentions specific interest in the protection of patient data (see SEC. 1177 of H.R. 3103, 1996). In a single, paragraph, the bill addresses the wrongful disclosure of individually identifiable health information, in large part, as it relates to insurance fraud and abuse.

> SEC. 1177. WRONGFUL DISCLOSURE OF INDIVIDUALLY IDENTIFIABLE HEALTH INFORMATION. "A person who knowingly and in violation of this part uses or causes to be used a unique health identifier; obtains individually identifiable health information relating to an individual; or discloses individually identifiable health information to another person, shall...be fined not more than $50,000, imprisoned not more than 1 year, or both; if the offense is committed under false pretenses, be fined not more than $100,000, imprisoned not more than 5 years, or both; and if the offense is committed with intent to sell, transfer, or use individually identifiable health information for commercial advantage, personal gain, or malicious harm, fined not more than $250,000, imprisoned not more than 10 years, or both."

This bill was the first step towards the development of a series of protections that would eventually become the HIPAA Privacy Law that we know today. However, much changed during the journey from the bill's initial proposal to passage of the final law and attendant guidelines—especially in terms of modifications made to the data privacy and de-identification standards. Early renditions of HIPAA provided very little guidance on how to define de-identified health information. Mass computerization of individual health information had only just begun, with electronic health records (EHR) making their first appearance in 1992. In the mid-1990s, with the rise of the internet and home computers, threats to data privacy elicited much fear within the American public (Best, Krueger, & Ladewig, 2006). Despite these concerns, when the bill went to congress in the summer of 1996, the disclosure of identifiable health information was not documented as a part of the discussion on the congressional record (Gingrich, 1996).

One year after its introduction, Latanya Sweeney, a computer scientist working at MIT, purchased a voter registration list for Cambridge, Massachusetts and cross-referenced that with a "de-identified" (meaning the names were missing but other information like birthdate remained) Massachusetts Group Insurance hospitalization dataset that was provided to researchers (Sweeney, 1997). Sweeney determined that by using birth date, gender, and 5-digit ZIP code she could match a patient's medical records with their name on the voter registration list. This meant that for only twenty dollars (the cost of the voter registration list), Sweeney could *potentially* identify (by name) some of the registered voters and their medical records which included sensitive information such as diagnoses, procedures, and medications. With this knowledge in hand, Sweeney famously mailed the governor his own medical records. This event fueled anxiety about the potential misuse of patient information and put data protection at the forefront of many conversations about privacy reform. Sweeney's 1997 study was central to the next chapter of the story of HIPAA's evolution—the 1999 Notice of Proposed Rulemaking (NPRM) (Standards for Privacy of Individually Identifiable Health Information, 1999; Barth-Jones, 2012).

In response to Sweeney's work, the 1999 NPRM proposed a very stringent definition of de-identified health information. Of particular interest to this paper is how the NPRM

defined the smallest unit of allowable geography as the state. All other geographic identifiers would be removed, meaning that street address, city, county, and both 3-digit and 5-digit ZIP were not permissible. This state-level geographic standard was too restrictive for any researcher interested in studying the geographic variation of health and disease such as geographers and epidemiologists. Under such rules, researchers would only be able to publish maps at the state-level (usually at the national extent). For most scholarship, this limit meant that only statistical point estimates (such as regression output) could be published under the safe harbor rule.

Fortunately for researchers, feedback from the 1999 NPRM's call for public comments pushed the HHS to allow some information about age and geographic area to be shared as de-identified information. The safe harbor standard's 3-digit zip code rule made its first appearance on a federal record (Standards for Privacy of Individually Identifiable Health Information, 2000). The rule states: "In the safe harbor, we explicitly allow...some geographic location information to be included in the deidentified information, but...zip codes must be removed or aggregated (in the form of most 3-digit zip codes) to include at least 20,000 people." Compared to the 1999 NPRM guidelines this safe harbor standard was much less stringent but still meant to withstand a population-level identification attack of the sort developed by Sweeny which required 5-digit ZIP codes to carry out.

This simple 3-digit zip code rule became more complicated in the decade after HIPAA was promulgated. The initial formulation seems clear (that 3-digit ZIP codes were the intended level of aggregation) however, subsequent modifications to HIPAA introduced ambiguity. Changes to the final rule in 2003 left out a key clause that made it clear that 3-digit zip codes would be the *only* permissible form of aggregation (other than the state-level). This contributed to the ever-growing ambiguity regarding the provision on geographic deidentification, and along with other nebulous aspects of the law, many researchers were finding it difficult to navigate HIPAA. As a result—with the passage of Health Information Technology for Economic and Clinical Health Act (HITECH) in

2009—the HHS was required "to issue guidance on methods for de-identification of protected health information (PHI) as designated in HIPAA's Privacy Rule." In response, the US Office of Civil Rights (OCR) held a workshop in 2010 to provide guidance on strategies for the de-identification of PHI. OCR used input from the panelists, including Latayna Sweeney and Daniel C. Barth-Jones (noted later in this paper), and workshop attendees to develop a lengthy guidance document (Office for Civil Rights, 2012). This comprehensive document is helpful in that it provides a more detailed description of the safe harbor rule, but unfortunately, it still contained the same ambiguous phrasing (regarding zip codes) found in the written law. To make matters worse, the landing page for the workshop on HIPAA's de-identification standard (which features the link to the guidance document page) refers to *geocodes* rather than ZIP codes (refer to Table 2 for full phrasing) which could easily lead readers to believe that any unit (not only zip codes) could be used for aggregation. These ambiguities, alongside inconsistencies in use and opinion found throughout the literature (explored below in section 3 below) about core HIPAA documents (e.g., HHS, 2003; Modifications to the HIPAA Privacy, 2013; OCR, 2012), may very well have contributed to the widespread confusion that continues today.

2.3 Why 20000 people?

Part of the ambiguity around using ZIP codes is tied to the 20,000-person threshold in defining safe harbor rules. The decision to allow sub-state level geographies, specifically ZIP codes, is partially tied to research on the role of population size in protecting privacy. In simple terms, by increasing the number of people reported within a given region, the chances of successfully matching an individual in that region to a record decreases. This is because the odds of a unique combination of identifying characteristics occurring in a population declines as the number of people in a dataset increases.

So how did the HHS determine that 20,000 was the appropriate population threshold? To answer this, we must look to the proposed final rule (Standards for Privacy of Individually Identifiable Health Information, 2000) as there is little to no discussion of

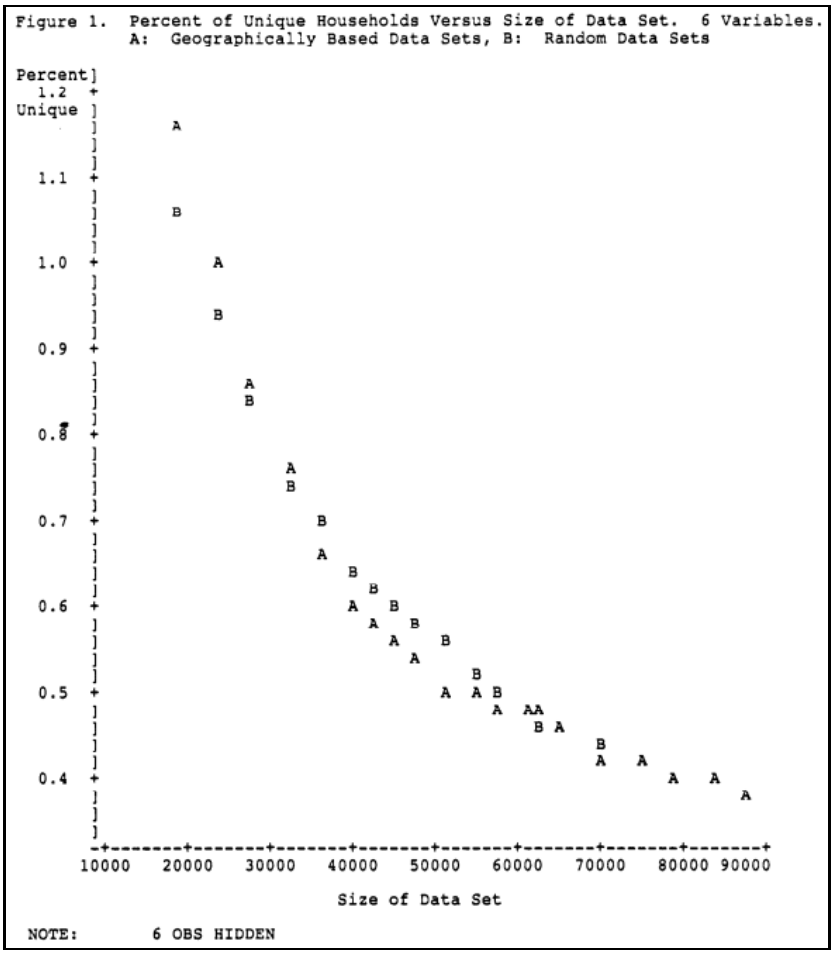
this determination within the literature or the HHS support and guidance webpages. In the final rule, the HHS points to the precedent of how the Bureau of the Census "shares geographical units only if they contain populations of at least 100,000 people" (The Federal Committee on Statistical Methodology, 1994). This standard is conservative and so the HHS turned to other sources to dropping the threshold lower (Standards for Privacy of Individually Identifiable Health Information, 2000).

The HHS drew on two simulation studies in particular, one by Greenberg and Voshell (1990) and the second by Horm (2000). These studies explored how the proportion of unique records within a dataset can be influenced by changes to the size of the population and the number and type of variables included. For instance, about 7.3% of records within the 1990 census are unique, or potentially identifiable, given the 100,000 person population threshold using standard census variables like age, race, ethnicity, sex, and housing/household information (Standards for Privacy of Individually Identifiable Health Information, 2000). But the proportion of unique records is a function of available information. Sharing a greater number of variables increases the potential to identify an individual, and for this reason, the Census Bureau population threshold increases from 100,000 to 250,000 or more when greater numbers of variables are released as microdata (The Federal Committee on Statistical Methodology, 1994).

However, there comes a point where increasing the size of the population no longer adds notable increases to data protection. In the case of census data, when only six demographic variables are shared, there is point of diminishing returns around about 20,000 people, per Figure 3.1 (Greenberg & Voshell, 1990). In addition to the number of demographic variables, the type of variables shared matters as well. For instance, a population of 25,000 contains 25% unique records when 9 variables were shared, but when the occupation variable is removed, this proportion drops to 10% (Horm, 2000). In this case, occupation can be a particularly identifying given that some occupations are

much rarer than others. The HHS drew on this scholarship to making their determination (Standards for Privacy of Individually Identifiable Health Information Final Rule, 2000):

> "After evaluating current practices and recognizing the expressed need for some geographic indicators in otherwise de-identified databases, we concluded that permitting geographic identifiers that define populations of greater than 20,000 individuals is an appropriate standard that balances privacy interests against desirable uses of de-identified data. In making this determination, we focused on the studies by the Bureau of Census cited above which seemed to indicate that a population size of 20,000 was an appropriate cut off if there were relatively few (6) demographic variables in the database. Our belief is that, after removing the required identifiers to meet the safe harbor standards, the number of demographic variables retained in the databases will be relatively small, so that it is appropriate to accept a relatively low number as a minimum geographic size."



3. 1 Plot of percent uniqueness according to the size of the dataset. This plot was used in the determination of the 20,000 population threshold (Greenberg & Voshell, 1990).

Additionally, the fact that HHS considers the 20,000 population stipulation the lowest bound could also be tied to adoption of the 3-digit ZIP. Although, 3-digit ZIP codes vary

widely in terms of the size of the population they contain (in 2020, ranging from 3,147 to 3,310,455 people), only 18 3-digit ZIP codes contained fewer than 20,000 people at the time safe harbor was first determined. Today, there are only 11 ZIP codes in the nation that are too small and would need to be merged with neighboring geographies to meet the minimum threshold of 20,000 people. Fortunately, because the majority of 3-digit ZIP codes contain populations well-over 20,000 people, researchers following the 3-digit ZIP code rule would not often be burdened with the task of data aggregation. Perhaps the HHS hoped that by using these 3-digit ZIP codes they could help enforce a more conservative following of the population threshold while also making the guidelines more straightforward. Unfortunately, this would not be the case in important ways.

3. Twin challenge #1: Ambiguity

The safe harbor rule seems straightforward when seen from the original 2000 final rule, but given the modifications, and how it appears in the literature today, it carries an essential ambiguity that has led to large gaps and disagreements in research and policy work. We first examine different interpretations of the rule based on these ambiguities and draw examples from scientific literature in order to show how different scholars rely on different interpretations. We then simplify the discussion by proposing that the crux of many disagreements — and the basis of productive ways forward — can be seen in terms of focusing on the use of 3-digit and 5-digit ZIP codes.

3.1 Safe-harbor provision and ZIP code ambiguity

The primary driver of disagreements in the literature seems to hinge on how individual researchers and teams interpret the role of ZIP codes vs. the 20,000 person threshold. This often comes to the fore in determining how much location data must be removed from patient data to satisfy HIPAA requirements.

The potential for misunderstanding stems from the one part of the provision—the piece regarding geographic information which states with respect to patient location data:

> (2) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
>
> (2a) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
>
> (2b) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.

Understanding of the HIPAA safe harbor rule has been furthered muddied by the different ways it is described by experts in the fields of public health and geography as well as by the guidance by HHS and OCR. Anyone reading the *background and context* section on the *2010 De-Identification Standard Workshop* page on the U.S. Department of Health & Human Services (HHS) website (OCR, 2017) could justifiably conclude that any aggregation of 20,000 people is in compliance with the safe harbor rule regardless of ZIP code. On the other hand, focusing on the ZIP code rules as they appear in the literature could lead someone to conclude that ZIP codes are the primary vehicle for data protection. This is because, in many cases, authors simply do not specify the type of ZIP code used in their work. This potential for ambiguity among different sources has likely contributed to the number of studies that have aggregated (or that have suggested the possibility of aggregating) in ways that do not align with the 2000 HIPAA final rule (Browne et al., 2014; Jung & El Emam, 2014; Mu et al., 2015; Acevedo-Garcia, 2001). Table 1 offers a number of different justifications for how scholars interpreted the safe harbor provisions.

Table 2. The various ways investigators interpret the geographic location stipulation of the HIPAA safe harbor rule.

Paper & Author	Interpretation
Confidentiality risks in fine scale aggregations of health data (Curtis et al., 2011).	"Unfortunately there are few guidelines with regards the release of aggregated data. A commonly discussed threshold between researchers is that health data should only be visualized for ZIP codes with a base population of no less than 20,000."
Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study (Sweeny et al., 2017).	"[T]he provision requires removing explicit identifiers (such as name, address and other personally identifiable information), reporting dates in years, and reducing some or all digits of a postal (or ZIP) code."
Workshop on the HIPAA privacy rule's de-identification standard (OCR, 2017).	"[The Safe Harbor approach] permits a covered entity to consider data to be de-identified if it removes 18 types of identifiers (e.g., names, dates, and geocodes on populations with less than 20,000 inhabitants) and has no actual knowledge that the remaining information could be used to identify an individual, either alone or in combination with other information."
Conforming to HIPAA regulations and compilation of research data (Clause et al., 2004).	"Implementation of these methods can be somewhat difficult for the clinical researcher for data sets of less than 20,000 records (as determined by collapsing populated geographic codes representing sparse populations)."
From Healthy Start to Hurricane Katrina: Using GIS to eliminate disparities in perinatal health (Curtis, 2008).	"The error of recording '70808' rather than '70806' in Baton Rouge would involve considerable changes in social, economic, and racial contexts. This is a problem if data are only available by zip code, which unfortunately is still too common in terms of releasing data for GIS analysis." "Although there are HIPPA regulations regarding the display of aggregate data on choropleth maps, these guidelines are generally considered too restrictive for useful cartography (only zip codes with more than 20 000 can be visualized)."
A linear programming model for preserving privacy when disclosing patient spatial information for secondary purposes (Jung and El Emam, 2014).	"A prevailing method to create de-identified data sets is to aggregate pre-defined areas, such as ZIP codes or counties, into a new area." "Yet, the first three digits of a ZIP code may be included, provided that at least 20,000 people share the same first three digits."
The Challenges of Creating a Gold Standard for De-identification Research (Browne et al., 2014).	"[The guidelines of the Privacy Rule] say that units smaller than a state should be redacted, although Baltimore has a population of well over 20,000, the size limit for Zip-Codes. D.C. was considered a state for this purpose."
Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation (Kayaalp et al., 2015).	"The Privacy Rule states that information about all geographic subdivisions smaller than state, except the first two digits of the zip code, must be de-identified. The third digit of the zip code can be left intact, only if the size of the population in the area of the censored two digits is greater than 20,000 according to the most recent census data."

Broken Promise of Privacy: Responding to the Surprising Failure of Anonymization (Paul Ohm, 2010).	"Id. § 164.514(b)(2)(B) (allowing only two digits for ZIP codes with 20,000 or fewer residents)."

The fact that a range of views exists is not surprising considering the ways in which HIPAA provisions have been interpreted within the fast-growing scholarly literature using spatial health data and among various online help resources. Understanding of the safe harbor provision is muddied by conflicting or ambiguous phrases that appear across a broad array of resources and how different scholars seem to follow different practices and procedures for handling patient location data. This profusion of differing practices, while perhaps engendering interesting conversation, likely comes at the cost of research outputs being unnecessarily overly masked in order to protect sensitive health data.

3.2 Two different interpretations

In order to find a way forward towards more standardized interpretations of HIPAA safe harbor rules, it helps to delineate two distinct ways of interpreting the safe harbor provision specific to location data (while recognizing that less-common interpretations may also exist). In essence, two different and competing interpretations have emerged: the 3-digit ZIP interpretation and the 5-digit ZIP interpretation.

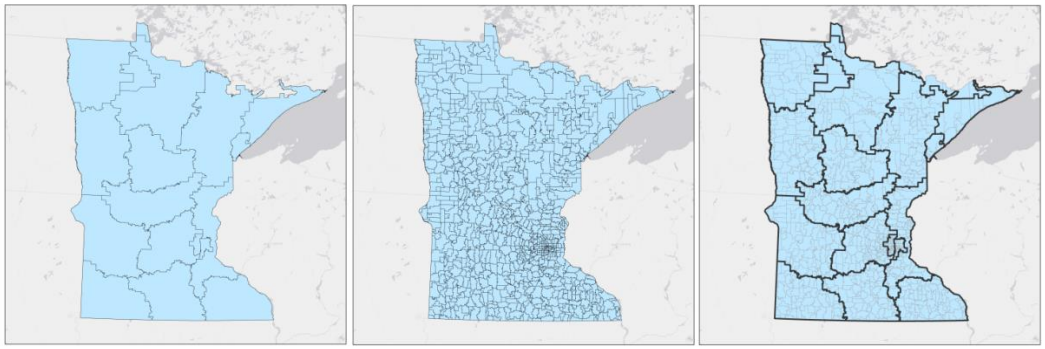
The 3-digit ZIP code interpretation. For many health researchers there is only one interpretation of the safe harbor provision. This is likely because much medical research involves working with data in its tabular form. For these investigators, a ZIP code is primarily a helpful 5-digit number that can be reduced to a 3-digit one. Consider, for example, an analyst receiving a spreadsheet of patient data from which to build her risk model. One column in the table would be designated for the location attribute (i.e., a column for ZIP codes). According to this rule, only the first three digits of the ZIP code are permitted to be shared (unless the population value is under 20,000 whereby the data is suppressed or converted to 000). For most lawyers, medical researchers, and anyone

using patient data in its tabular format, there is little ambiguity in the safe harbor standard.

The 5-digit ZIP code interpretation. To those who see ZIP code data primarily as spatial data, the privacy rule elicits some confusion. While a ZIP code is a 5-digit number, to geographers and a growing number of other scholars who use spatial data, a ZIP code is also an area on a map. ZIP codes divide regions into smaller areas designed to aid post-delivery. There are both 3-digit ZIP code areas (Figure 3.2) and 5-digit ZIP code areas (Figure 3.3). Five-digit ZIP codes areas are nested within 3-digit ZIP code areas (Figure 3.4). People who work with spatial data are likely familiar with this hierarchy of spatially nesting areas and how it can lead to conflicting interpretations of provision §164.514(b)(2a) which states:

> (2a) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people;

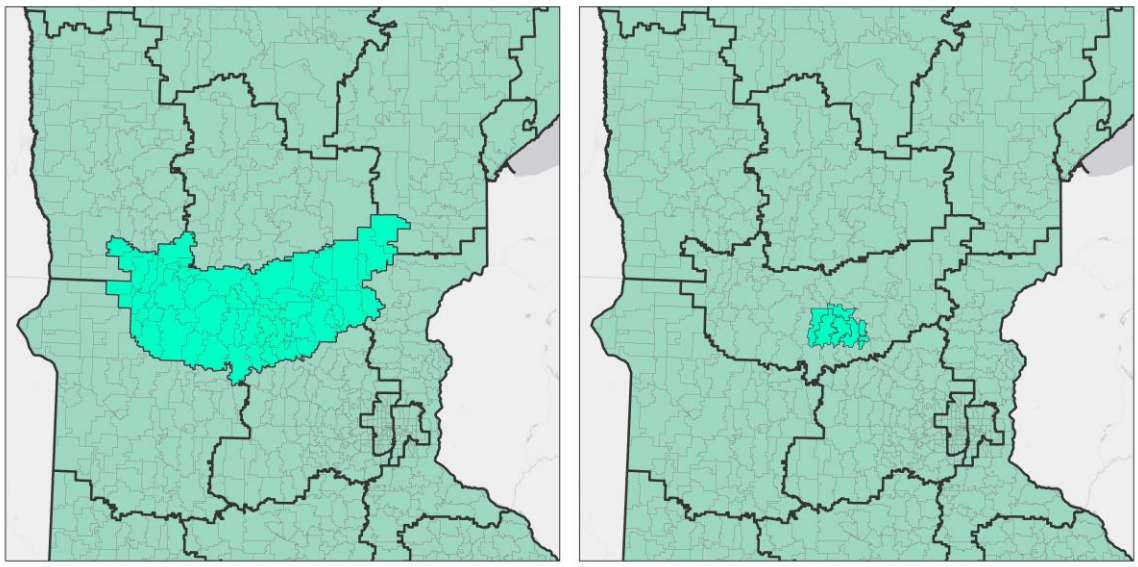
In this view, there are two ways of reading "ZIP codes with the same three initial digits", namely either: 1) 3-digit ZIP codes (as described in the previous paragraph) or 2) 5-digit ZIP codes that share the same three initial digits.



3. 2 Three-digit Zip code boundaries. 3. 3 Five-digit Zip code boundaries. 3. 4 Five-digit Zip codes nested within three-digit Zip codes.

The root of this apparent ambiguity comes from the term "all ZIP codes." If we interpret "all ZIP codes" as "all of the 5-digit ZIP codes", then the 3-digit ZIP code rule would still apply because when you combine all of the 5-digit ZIP codes together you are left with a 3-digit ZIP code area (Figure 3.5A). If however, "all ZIP codes" were interpreted as "all

5-digit ZIP codes within the aggregation", a less conservative interpretation emerges where 5-digit ZIP codes can be combined to meet the 20,000 population threshold as long as *all* of the 5-digit ZIP codes used have the same three initial digits (Figure 3.5B). Simply put, this interpretation would permit investigators to aggregate 5-digit ZIP codes when they all fall within the same 3-digit ZIP code area. The large difference in areas highlighted by Figures 3.2 and 3.3 demonstrates the impact of these two competing interpretations.

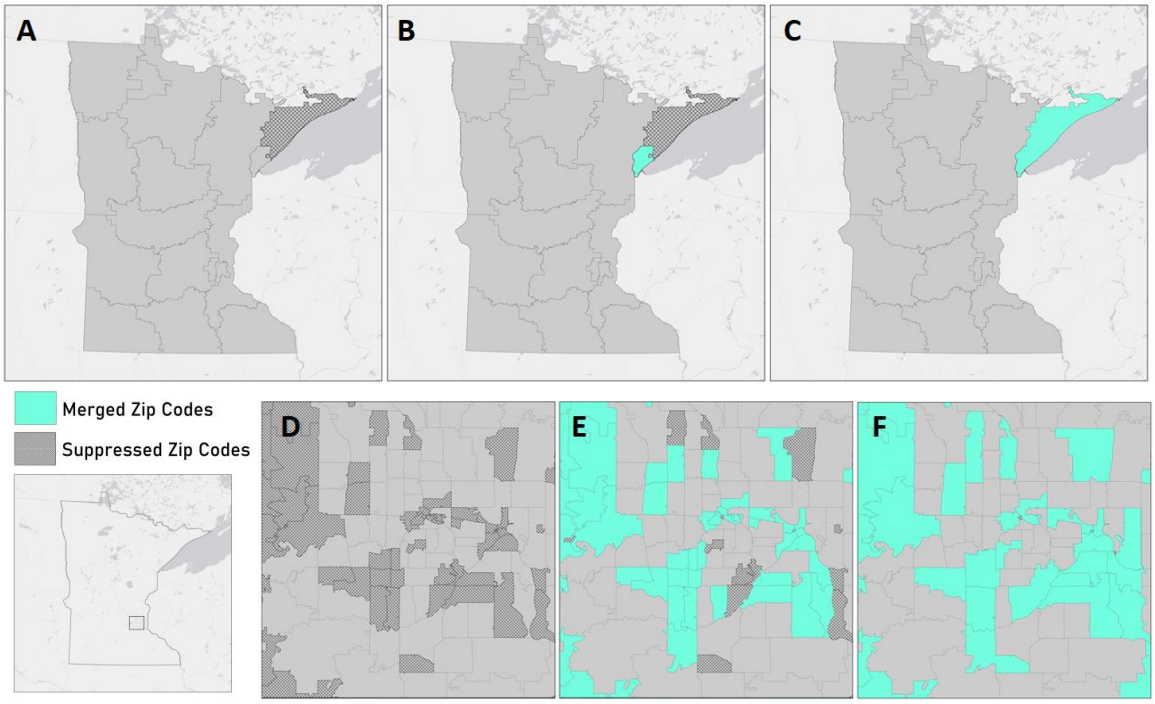


3. 5 A) All of the 5-digit Zip codes beginning in "563". B) An aggregation of 5-digit Zip codes that all begin with "563" that contains more than 20,000 people.

3.3 Drivers and implications of the two interpretations

Comparing studies that use 3-digit vs. 5-digit ZIP codes illuminates a potential cause for the existence of competing interpretations tied to whether the work uses tabular data or spatial data. In the case of either the 3 or 5-digit ZIP code interpretation, the tabular data can appear in essentially the same format (only containing the first 3 digits of a ZIP code). These same data mapped, however, would be **very** different. A researcher operating under the 3-digit interpretation would share maps of patient data at the 3-digit ZIP code level (Figure 3.6A), and if a 3-digit ZIP code contained fewer than 20,000

people it would be merged with a neighboring unit (Figure 3.6B and 3.6C). The corresponding tabular data for these maps would only contain three-digit ZIP codes. However, investigators operating under the 5-digit ZIP code interpretation could share maps at the 5-digit ZIP code level (Figure 3.6D), and if the 5-digit ZIP code contained less than 20,000 people it would be merged with neighboring units that share the same first initial digits (Figure 3.6E and 3.6F). The corresponding tabular data for these maps would only contain the first 3-digits of a ZIP code as well, however since more than one aggregation would fall within each 3-digit ZIP code area, there would be multiple records with the same 3-digit ZIP code.



3. 6 The aggregation process as see within 3-digit Zip codes (top row) and 5-digit Zip codes (bottom row). Zip codes with populations less than 20,000 people are suppressed. To address suppression, low-population Zip codes are merged with neighboring Zip codes to meet HIPAA requirements. It is not in adherence with HIPAA safe harbor to use 5-digit Zip codes as the unit of aggregation.

These differences are not hypothetical because relevant examples are abundant within the literature. Bearing in mind that researchers rarely describe their decision making in detail, there is a body of work that seems to operate under the 3-digit ZIP code interpretation (e.g., Barth-Jones, 2012; Browne, Kayaalp, Dodd, Sagan, & McDonald, 2014; Janmey &

Elkin, 2018; Malin, Benitez, & Masys, 2010; Nicholson & Smith, 2007; Sweeney et al., 2017; Tellman et al., 2010). There is another realm of scholarship that appears to operate under the 5-digit ZIP code interpretation (e.g., Curtis, 2008; Curtis, Mills, Agustin, & Cockburn, 2011; Wang, Guo, & McLafferty, 2012; Acevedo-Garcia et al., 2001), as well as related work that seems to suggest the capability of aggregating any geocode to meet the 20,000 threshold (e.g., Browne et al., 2014; Jung & El Emam, 2014; Mu et al., 2015). These are some of many potential examples of how there appears to be a divide between the 3-digit and 5-digit ZIP code interpretations of HIPAA.

Interestingly, there appears to be some commonality within and differences between disciplines in regards to the way safe harbor is interpreted. While this paper does not attempt to do a full literature review, anecdotally, of those studies cited in the paragraph above, all those operating under the 3-digit ZIP code interpretation are authored by epidemiologists, medical researchers, or computer and information scientists, while the papers backing the 5-digit ZIP code interpretation are authored by geographers. Although this is just a sample of a larger literature, there seems to be a trend where spatially-oriented researchers are more likely to embrace the 5-digit interpretation or a more lenient understanding of the rules around a threshold of 20,000 people. This is not surprising given that geographical research often necessitates a map, and three-digit ZIP codes are not intuitive map units. It is also the case that 3-digit ZIP codes are not easy to find in the form of shapefiles, or mapping files, that are often used for research. Neither census.gov nor USGS offer data at the 3-digit ZIP code level. In fact, at the time of writing, we can only find two sources that provide data for download in the form of 3-digit ZIP code boundaries for the U.S. and both of these sources are proprietary (Esri's ArcGIS Online and Caliper's Maptitude). Even without having access to these proprietary resources, it *is* possible to create these boundaries on your own. However one would think that, since 3-digit ZIP codes are the required units for display under HIPAA law, they should be more readily available online. On the other hand, data at the 5-digit zip code level is easy to find online and appears abundantly within the public health literature. The extent to which the dearth of 3-digit ZIP code

map data plays a role in the misunderstanding of the safe harbor rule is unclear, but one can't help but wonder whether the widespread confusion would exist if 3-digit ZIP code mapping files were available for download on the HHS website.

The potential implications of misunderstanding the privacy guidelines are profound when considering that researchers share patient data in inconsistent ways that bear on both efficacy of health interventions and potential for privacy breaches. When studies share aggregate patient data at the level of the 3-digit ZIP code their output is generally not useful for identifying local distributions of health and disease, although they do provide a more generous degree of data security. When studies share PHI at the 5-digit ZIP code level, they can provide a much more useful depiction of the spatial heath dynamics at hand, but at the cost of weaker data privacy.

In terms of this tradeoff, the difference in identification risk between 3-digit and 5-digit ZIP codes is substantial enough to warrant alarm, as discussed in detail in the next section (Sweeney, 1997). At the same time, the difference in spatial resolution between the two forms of ZIP codes carries its own and potentially problematic costs. For instance, one study demonstrated how different disease patterns emerge depending on whether 3-digit or 5-digit ZIP codes areas are used and, with an example dataset, the authors showed that if 3-digit ZIP codes areas are used to determine how to best distribute N95 respirators during a pandemic, it would result in a surplus of supplies for healthcare workers in some communities and shortages others (Tellman et al., 2010).

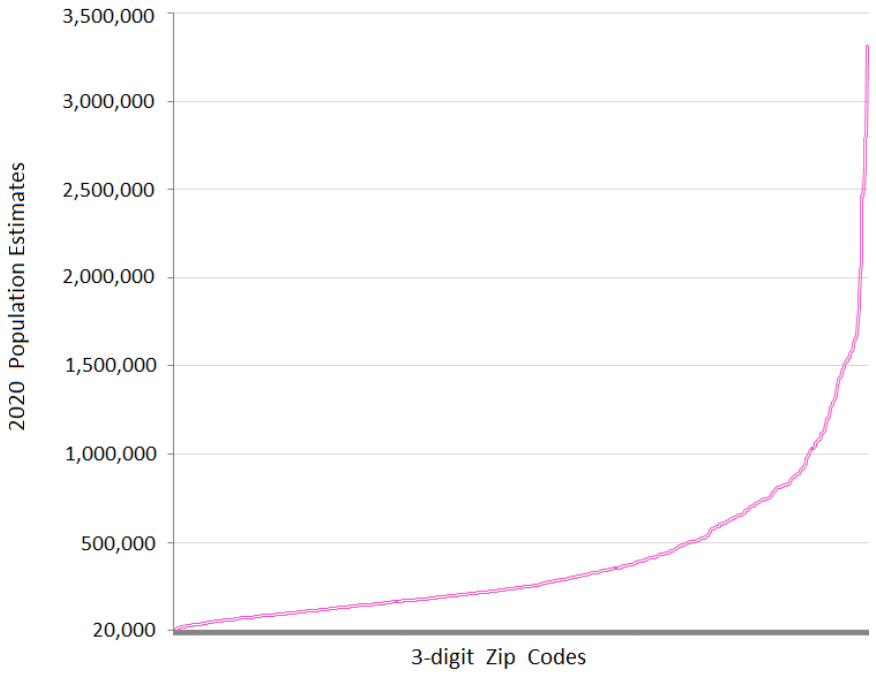
4. Twin challenge #2: Data loss

Even after gaining a clearer understanding of HIPAA law and how it is meant to be interpreted, one more challenge remains— namely that HIPAA guidelines are very likely too strict in general resulting in an unnecessary large degree of data loss. The following sections provide insight into the extent of the data loss that takes place when adhering to

HIPAA Safe Harbor's 3-digit ZIP code rule and how other (non-HIPAA compliant) interpretations can reduce data loss while not add much in terms of privacy risk depending on the kinds and amount of data being shared.

4.1 Data loss from 3-digit ZIP codes & 20,000 people

Opting for the 3-digit ZIP code interpretation is a conservative choice that has a number of negative implications for research and policy. The 3-digit ZIP code interpretation is very cautious with respect to adhering to the 20,000 person rule. Bear in mind that, as of 2020, the average population contained within a 3-digit ZIP code is 397,372 people, which is almost four times the population threshold of 100,000 required by the Bureau of the census for the release of microdata (individual response data from the census). Thirty years after the initial rule, there are now only eleven 3-digit ZIP codes that require suppression (because they have fewer than 20,000 people within them). The number of ideal units containing small, yet acceptable, populations is disappointingly low—only 12 units contain between 20,000 - 30,000 people and only 14 contain between 30,000 - 40,000 people. Just over 92% percent of 3-digit ZIP code geographies contain more than 60,000 people, or at least three times the 20,000 threshold. In simple terms, we should expect that most geographies shared under the 3-digit ZIP code safe harbor standard will contain populations far greater than the 20,000 threshold.



3. 7 Three-digit Zip codes (100-999) ordered least to greatest by population from 2020 estimates from the ACS.

Given that most 3-digit Zip code geographies contain well over 20,000 people, under the HIPAA safe harbor provision, the majority will have a very small proportion of uniques. However, a few places will have a proportion of unique records considered to be relatively more risky in terms of patient protection. In any case, the small number of instances that contain the "riskier" low-level minimum populations still meet the minimum acceptable level of risk (which if we look back at Horm's simulation study, we can estimate this to be a little over 10% proportion unique). This is a little bit higher than the 7.3% estimated uniques in the 1990 census microdata, but the HHS points out that the actual risk will be much lower because of the limited number of publicly available tables that can be used to compare the patient data with. Here, it is also important to recall that these risk estimates are also subject to the previously mentioned myth of the perfect population register. Finally, HHS suggests that the relatively low probability of success should be a deterrent in and of itself.

One interpretation of this threshold is that, if the HHS is okay with some units being shared at the level of 20,000 people, could all units be shared at that resolution? After all, if populations of 20,000 meet the minimum acceptable level of risk, then what's stopping investigators from aggregating 5-digit ZIP codes to meet this requirement? 3-digit ZIP codes are rather impractical for research purpose and so it is very uncommon to find a map shared at this level. For this reason, it is easy to see how researchers could come to believe that the 5-digit interpretation is permissible if they haven't given the legal documents a thorough read.

Aggregating 5-digit ZIP codes to create the finest-grained units possible that also still meet the 20,000 person threshold is tempting, because this would allow investigators to meet the minimum acceptable level of risk in a way that enables the sharing of maps with more detailed and consistent geographies than that provided by 3-digit ZIP codes. In this scenario, there would be slightly greater risk of identification due to the minimum population size, but it would still seem to be an acceptable level of risk as long as the 18 other safe harbor restricted identifiers were removed. The problem that remains is that one of the 18 identifiers isn't being *fully* removed in this scenario. By aggregating 5-digit ZIP codes, an individual record contains more information than a single 3-digit ZIP code—it now also contains a handful of 5-digit ZIP codes that could be used to further narrow down the possible matches. For this reason, 5-digit ZIP code aggregations do not meet HIPAA safe harbor standards.

However, depending on what other information is kept, it is reasonable to believe that sharing a map of patient data, stripped of age and other demographics, at the aggregated 5-digit ZIP code level would lead to a very low (certainly quite low) risk of identification. One study showed that certain elements from the list of 18 identifiers can still be shared without jeopardizing patient privacy "when other features are reduced in granularity". Specifically, Malin and colleagues found that more detailed age data (beyond what is permitted by safe harbor) could be shared when they coarsened the

specificity of other variables such as ethnicity (Malin et al., 2011). The authors noted that every dataset is different and, because of this, alternative de-identification practices can be used to enable the safe disclosure of patient data that is normally suppressed under the safe harbor method. This means that there is potential for 5-digit ZIP code information to be safely shared in aggregated form as long as other identifying information is suppressed.

In sum, it may be time to rethink the one-size-fits all strategy that is the safe harbor method. It is reasonable to ask whether aggregating 5-digit ZIP codes to regions that contain at least 20,000 people could achieve a "sufficiently low" risk of identification when other patient information is suppressed such as date of birth and gender. It would be even more reasonable to suggest that aggregating 5-digit ZIP codes could work if no patient information was shared other than diagnosis and location. Andrew Curtis and colleagues tested this such claim in a study that found that, when put to the test, students were unable to identify individuals in simulated cancer maps (Curtis et al., 2011). There was little reengineering risk even at aggregated resolutions finer than 20,000 people. Up to this point, this paper has pointed out the ambiguities within the safe harbor standard while shedding light on some of the arbitrary determinations made by the HHS that have contributed to a perhaps overly conservative definition of privacy. The following section takes a closer look at how the safe harbor rule has been criticized for being too stringent and, at the same time, not protective enough, specifically when it comes to identification risk.

4.2 Do the privacy gains justify the amount of data loss?

In order to dive deeper, we must go back and consider the influence of Sweeney's 1997 population-level identification attack. As stated previously, this initially resulted in the decision to bar both 3-digit and 5-digit ZIP codes from de-identified data, but after taking public comments, HHS reconsidered and 3-digit ZIP codes were deemed permissible as long as they contained a population of at least 20,000 people. HHS justified their

restrictions by citing particular studies which led them to believe that the combination of 5-digit ZIP code, gender, and date of birth (DoB) would be enough to potentially identify a great deal (more than half) of the U.S. population on the basis of uniqueness (L Sweeney, 2000). Note that, to be considered "unique", a record must contain a combination of characteristics that make it different from all other records in that table (Zayatz, 1992). If the number of unique individuals within the U.S. population was really as large as Sweeney reported it to be, the motion to block 5-digit ZIP code and DoB under safe harbor seems quite justified. However, some have pointed out that the combination of these three identifiers—even with their formidable discernibility capabilities—might not be as threatening as Sweeney's article makes it out to be.

Daniel C. Barth-Jones describes the "myth of the perfect population register" in his 2012 paper, which points out how many investigators often forget to account for the people missing from the lists used to link individuals to their medical records. These missing populations add significant uncertainty into the calculation of true population uniqueness (Barth-Jones, 2012). For this reason, the actual proportion of unique individuals on a list cannot be determined with 100% certainty if potential matches exist off the list. Therefore these kinds of studies must be careful in the statements they make—oftentimes including phrases such as "likely unique" or "potentially identifying" as certain identification cannot be claimed without a list of the entire population or the knowledge that the individual under identification attack was indeed contained within both lists.

Consider for instance, Sweeney's 1997 paper which the 1999 NPRM cites saying "A 1997 MIT study showed that, because of the public availability of the Cambridge, Massachusetts voting list, 97 percent of the individuals in Cambridge whose data appeared in a data base which contained only their nine digit ZIP code and DoB could be identified with certainty" (Standards for Privacy of Individually Identifiable Health Information, 1999). According to this, nearly all of Cambridge voters can be identified

using the combination of date-of-birth and 9-digit ZIP code. Within Sweeney's paper, she states that this proportion of people can be "uniquely identified" on this basis, however, these individuals are only uniquely identifiable within the population of registered voters and not within the general Cambridge population (see Barth-Jones for full explanation). This means that, in order for an intruder to identify an individual's medical record, they would have to know that the individual exists on both lists AND that no other person in Cambridge shares the same DoB and 9-digit ZIP code. When deciphering the data, the intruder must account for 35,000 non-registered voting-aged people living in the city—any one of which could be the true subject of the medical record of interest. Unaccounted for populations inject much uncertainty into the identification of unique records (in the case of Sweeney's 1997 study 35% error). With an imperfect population register, as exemplified in the Cambridge attack, an intruder would be able to identify with 100% certainty no one. Barth-Jones concludes that the governor was likely only identifiable based on the fact that he was a public figure who had a public hospitalization. The date of hospitalization was known as well as his DoB, gender, and ZIP code; moreover it could be easily assumed that he would be a registered voter. In instances such as this (having information a priori)[1], an intruder can be confident in the unique match.

It is unclear whether the HHS wrote the NPRM with a full understanding of methodological limitations of voter list-based identity attacks of the kind described by Barth-Jones. It is possible that the clause "...could be identified with certainty" was taken without really considering the implications of the prior clause "...whose data appeared in the data base". Many assumptions need to be met before we can ignore the myth of the perfect population register. In this example, in order to identify 97% of the individuals with certainty, we would need to be sure that none of the 54,805 voters on the voter list

[1]The safe harbor law has an additional stipulation (the very last line of the provision) which was built in to protect against identification attacks targeting highly identifiable people (like the governor). This stipulation reads: "The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information."

had the same birthdate as a non-voter living in their neighborhood. We might then wonder how would the identifiable 97% on the list compare to the proportion identifiable in the entire Cambridge population? This is something we can't determine because we don't have a population register, but given that the total population of Cambridge is approximately 88,000 (Barth-Jones, 2012), there is quite a bit of room for error. If the HHS based their development of safe harbor on a limited understanding of these complexities, it might lead us to wonder whether the level of protection delineated within the safe harbor standard is overly conservative.

Nevertheless, even if the HHS misunderstood how Sweeney was using the term "identifiable" within her 1997 paper, there is still room for concern about how far to read into the study. Sweeney's work is bold, insightful, and conveys a critical message: our private information is vulnerable to attack. What's unclear is the extent to which we *understand* the vulnerability. Even with the injection of uncertainty from missing populations, the risk for identification may still be considered too high and the implications would be quite serious. Let's go back to Barth-Jones' review of Sweeney's 1997 attack, which finds that somewhat fewer (but perhaps not much fewer) than 29,000 people out of 88,000 in Cambridge are identifiable (if the record is unique and the data intruder already knows that the individual is on both lists). Depending on the motive of a data intruder, this might not be that far from likely. Consider that it is easier to link a specific person to their medical record than it is to link a specific medical record to the person it belongs to. This is because a motivated attacker would have likely collected background information on the person a priori. The data intruder likely has a target in mind—someone that they know—and therefore it is not that unlikely for them to already have information on the target's voting behaviors and place of work—allowing the intruder to determine the employment insurance coverage that could be used to confirm the target's presence on the insurance hospitalization data list. Moreover, even without knowing with certainty if the target of the attack is on both lists, the fact that the chance of a false positive (matching a record to a voter on the list when the record actually belongs to a non-registered voter) occurring could be perceived as highly unlikely by the

attacker—which could encourage them to carry on with their plans regardless of the potential false positive.

The combination of DoB, gender, and 5-digit ZIP code can be troubling when shared in conjunction. The question that remains is: Can this combination of identifiers be reworked to reduce the risk for identification? Within the literature on microdata anonymity, ZIP code, gender, and DoB are actually not considered full identifiers themselves, but rather, they are quasi-identifiers that can be used in combination to find unique instances. The term "identifier" is reserved for information that uniquely identifies an individual such as a Social Security Number (Ciriani, De Capitani di Vimercati, Foresti, & Samarati, 2007). Nevertheless, quasi-identifiers can be dangerous when used in combination, but how dangerous are they? In order to gain some insight into this question, we must look more closely at how identification risk has appeared within literature relying on the HIPAA safe harbor method.

4.3 What level of data loss defines sufficient data protection?

What is an acceptable level of identification risk? There is no universally recognized standard that defines what a sufficient proportion of unique records should be. Some have suggested that the nationally accepted standard of re-identification risk is defined by HIPAA's safe harbor standard itself (Janmey & Elkin, 2018), but recall that the safe harbor standard was derived somewhat arbitrarily, being loosely based on rules used by the Bureau of the Census and a couple simulation studies. In fact, when determining the population requirement of the HIPAA safe harbor rule, the HHS made the following statement in regards to defining "minimal risk":

> With respect to how we might clarify the requirement to achieve a "low probability" that information could be identified, the Statistical Policy Working Paper 22 referenced [in the 2000 final rule] discusses the attempts of several researchers to define mathematical measures of disclosure risk only to conclude that "more research into defining a computable measure of risk is necessary." When we considered whether we could specify a maximum level of risk of disclosure with some precision (such as a probability or risk of identification of

<0.01), we concluded that it is premature to assign mathematical precision to the "art" of de-identification.

Because twenty years later there is still no threshold defining “sufficiently low probability,” investigators fall back on the safe harbor standard as a point of reference for comparing different levels of data protection. De-identification with the safe harbor method is said to leave somewhere around 0.03% or 0.04% records within the US population vulnerable to identification (NCVHS, 2007; Barth-Jones, 2012), but this proportion fluctuates according to the geographical extent of the dataset, where some regions have much smaller proportions of unique records and others have much higher. Specifically, re-identification risk has been found to range from 0.01% to 0.19% (Malin et al., 2011), 0.01% to 0.25% (Benitez & Malin, 2010), and 0.013% to 0.22% (Kwok, Davern, Hair, & Lafky, 2011) on a state by state basis.

Most studies estimate the identification risk under safe harbor to be rather low. Despite this, there is no consensus on whether or not safe harbor standards are sufficient for protecting patient data. In other words, “sufficiently de-identified” is subjective and, on occasion, very similar proportions of unique records have evoked very different assessments. For example, Sweeney asserts that her estimated safe harbor re-identification risk of 0.04% of the US population is not a sufficient privacy guard (NCVHS, 2007; Sweeney, 2017) while Barth-Jones suggested that the risk would actually be less than 0.03% (when using a voter list attack strategy), and that this proportion is in fact sufficient, going on to compare the identification risk under safe harbor to the likelihood of being struck by lightning (Barth-Jones, 2012). A re-identification attack by Kwok and colleagues re-identified only 2 of 15,000 individuals (0.013%) from a safe harbor protected dataset and the intruder was provided with a substantial amount of information from a market research company (Kwok et al., 2011). Kwok et al concluded that there was a low risk of re-identification and that masking with safe harbor makes re-identification a challenging task. Others have asserted that safe harbor is too stringent. Bradley Malin suggested in a 2011 article that

the safe harbor method was too conservative because it is possible to release more detailed information without presenting greater risk than that provided by the safe harbor method. On the other hand, a 2016 study found that even when data seems sufficiently masked, computer science models can be used to identify a large proportion (42.8%) of patients by linking demographics such as age, sex, hospital, and year (O'Neill L, Dexter F, 2016). Although specific to a single case study, this is an a high and very likely unacceptable level of risk! More recently, Janmey and Elkin suggested that the safe harbor standard is sufficient for preserving privacy at an overall population level (2018). However, they also found that encounter notes within data can sometimes include indirect identifiers that can be used to help match records, and this could increase the risk of identification to 0.07% which does not meet the safe harbor criteria of sufficiently de-identified.

It is safe to say there is disagreement about what is sufficient in terms of data protection. This type of risk calculation is complicated in and of itself and a concept like 'sufficiency' is necessarily a judgement call. Recall that identification risk depends not only on how the data is released, it also depends on the alternative lists publicly available to the data intruder. Sweeney described how identification risk for safe harbor abiding datasets can be as high as 25% when the intruder uses more than just a voter registration list (Sweeney et al., 2017). Other detailed registries can be used to re-identify masked data such as real estate tax data, credit reports, and property records. Moreover, identification risk can foreseeably jump much higher—far beyond the expected ranges—for certain areas where the demographics of the base populations allow an intruder to easily narrow down potential matches based on age or ethnicity, as seen in regions dominated by college dorms, ethnic enclaves, or transient communities (Sweeney, 1997; O'Neill et al., 2016). Sufficient data protection (leaving aside the definition of sufficiency) will always be dependent on the dataset being masked because a slew of factors determine the overall identification risk.

5. Ways forward

So far we have focused on the two key issues of safe harbor provisions — the confusion around which ZIP codes to use and whether the rule warrants an unnecessarily large amount of data loss. Reviewing the process by which the safe harbor concept came into being provides insight into the intended interpretation of the provision and the motivations that guided its development, but it is a first step. The ambiguity about how to best interpret and use ZIP codes or other geographic identifiers persists and there is no clear consensus on what defines sufficient minimal risk. Here we explore new approaches to data privacy and how they may meet the needs of some researchers, but we conclude by arguing that the most promising way forward to addressing the twin problems of safe harbor is to steer away from one-size-fits-all guidelines and towards deeper assessments of domain-specific and data-specific modes of masking that could offer a middle ground between useful data and protected data.

5.1 New approaches to de-identification

In the face of the complex nature of re-identification risk, scholars and policy makers have begun to advocate for the widespread adoption of k-anonymity or differential privacy methods (Sweeney et al., 2017). The primary argument for these approaches is that de-identification methods should come with privacy guarantees, especially as technology advances and powerful automated systems can be made to search for matches between multiple public lists. For this reason, although k-anonymity and differential privacy cannot necessarily guarantee data security, these methods have been getting much attention as of recent because they provide a sort of privacy guarantee that offers more complete data protection than the traditional masking approaches.

K-anonymity ensures that no unique records exist in the dataset and further requires that each record has a minimum of "k-1" common records (those that have the same quasi-identifiers) so that they can't be differentiated and therefore identified with certainty (Samarati & Sweeney, 1998). K-anonymity can be achieved through many traditional methods such as through jittering, aggregation, and location swapping, and often provides

a higher level of protection than if one were to use one of these traditional methods alone. Despite this, k-anonymity is not impervious to intruder attacks. An intruder can still use background knowledge to narrow down the possible matches to increase the likelihood of identification such as in the case of a homogeneity attack (attacks based on data that contain identical values for an attribute) in which a region with a homogeneous population containing similar values for a record in the table can be used (alone or linked with other data) to identify an individual or diagnosis. Therefore, k-anonymity, strictly speaking, does not guarantee privacy. However, it guarantees non-uniqueness which, in the absence of outside knowledge, provides considerable data protection and, for this reason, k-anonymity remains a popular approach.

Differential privacy (DP) is attracting attention as a newer approach to protecting sensitive data that assures a very low likelihood of individual identification. The most common definition of differential privacy is that of epsilon (ε) differential privacy introduced by Cynthia Dwork and colleagues (2006). Dwork's ε-differential privacy involves creating a synthetic aggregated dataset from an original unprotected dataset which ensures that an individual record cannot be identified. This simulated data is built by injecting a predetermined amount of noise (based on a Laplace distribution) into the original aggregate table in a way that does not significantly influence the output (of queries into particular pre-specified relationships). In other words, the aggregate table is systematically adjusted in a way that secures individual privacy while also ensuring that the data provides similar results to what would have been given if the original data was used in a pre-specified analytical model. The way in which this is achieved also makes it so that if any one individual was removed from the dataset, it would not influence the overall results. This means that epsilon differential privacy provides relative guarantees about disclosure risk, and essentially promises that "…any given disclosure will be, within a small multiplicative factor, just as likely whether or not the individual participates in the database." (Dwork, 2006).

Unlike k-anonymity, differential privacy protects data under the assumption that an intruder has close to perfect knowledge and, in doing so, differential privacy offers a level of protection unlike others. Differential privacy does not succumb to the same weaknesses of traditional methods (including the homogeneity attack), and provides stronger data protection against differencing, linkage, and reconstruction attacks (Dwork & Roth, 2014). Additionally, due to its robustness, differential privacy has the advantage of reducing improper data analysis techniques by limiting the ability of a single observation to have an effect on the result, which helps to deter things like p-hacking, HARKing, and overfitting models (Dwork et al., 2015). For these, and many other reasons, differential privacy has gained much attention over the past two decades. In fact, differential privacy methods have the potential to replace existing masking methods and have already been adopted by Apple and the Bureau of the Census—which intends to use differential privacy to protect the 2020 census microdata. Differential privacy is not infallible; it offers "an extremely strong guarantee, it does not promise unconditional freedom from harm" (Dwork & Roth, 2014).

Because differential privacy provides a higher level of protection than many other methods, it potentially offers a way for researchers to share data at more detailed levels than previously allowed under safe harbor. Consider the example of disease surveillance mapping. Safe harbor's minimum population requirement of 20,000 people is rather limiting in terms of map resolution. A map with units that contain 20,000 people would not provide enough detail to be helpful to researchers, policy makers, or community members. Differential privacy, however, would allow investigators to share maps at much finer scales (down to the neighborhood-level) without putting patients' identities at risk.

So why not use differential privacy? Because it has critical drawbacks for research use (Muralidhar et al., 2020). For instance, a map created from a differentially private aggregated table displays simulated data, so it is possible that some regions on the map

would not accurately reflect the original data—especially at finer scales where the population numbers are lower. Santos-Lozada and colleagues found that the infusion of noise from DP methods impacts observed distributions differently for different demographics, meaning that DP has the potential to bias understandings of health disparities at the national level (2020). In particular, the authors demonstrated how mapping differentially private data led to "overestimates of population-level health metrics of minority populations in smaller areas and underestimates of mortality levels in more populated ones", and these effects were dramatic. For instance:

> "…in McCulloch County, Texas, the mortality rate ratio for non-Hispanic blacks is 75.9, indicating the mortality rate would be 24% lower under the current methodology compared with the differential privacy methodology. Similarly, in Clarke County, Virginia, the mortality rate ratio for Hispanics is 121.4, indicating the mortality rate would be 21% higher under the current methodology compared with the differential privacy methodology. At the same time, the non-Hispanic white mortality rate ratios were essentially unchanged for these two counties, at 100.3 and 99.8, respectively, meaning substantial biases may enter into understandings of disparities."

The implications of differential privacy for research are dire and the recent move by the Bureau of the Census to adopt this approach for the 2020 census microdata has drawn much attention to its advantages and disadvantages (Oberski & Kreuter, 2020; Ruggles, Fitch, Magnuson, & Schroeder, 2019). Census data is one of the largest sources of sociodemographic data used by social scientists and therefore, differentially private methods threaten to degrade the reliability and effectivity of social science research. Other than threats to data accuracy and biases, another source of concern regarding 2020 census data is that these differentially private tables would not enable exploratory data analysis. This is because differentially private data is synthetic and therefore relationships cannot be explored unless they were pre-specified when the synthetic table was created. For this, it is very likely for differential privacy to interfere with the process of data-driven scientific research, pushing some scholars to suggest that perhaps "…differential privacy goes far beyond what is necessary to keep data safe" (Ruggles et al., 2019).

There is much uncertainty in regards to the practicality of differential privacy for the protection of large-scale, sensitive data. Differential privacy is a relatively new concept for many social scientists and epidemiologists. There is a dearth of investigations into differential privacy within the social science literatures, and particularly in regard to the impact it might have on health mapping—we could only find only one study at the time of writing (Santos-Lozada et al., 2020) but expected more given the attention giving to differential privacy and many unanswered questions it poses. What are the implications of differential privacy on mapping in terms of accuracy and use? How do differentially private maps compare to maps of original raw data? Furthermore, it is unclear how differential privacy stands within institutional IRBs. This is relatively new territory and it is likely that many HIPAA compliance officers are not familiar with differential privacy. As part of our examination of the history of HIPAA, we spoke with legal experts and HIPAA compliance officers. One such officer, being introduced to differential privacy, stated that "this doesn't play into our office's considerations of deidentification." Differential privacy holds some promise for mapping spatial data but at known and unknown costs.

5.2 Current state and future research

Despite ongoing interest in expanding use and sharing of health data mapping, the safe harbor rule stands as the primary guidance for those interested in sharing maps. It is far from perfect in that for many scholars, it is ambiguous and either too stringent or not sufficient in terms of securing data or lessening data loss. Alternative methods exist that have the potential to do a better job but they come with their own drawbacks. HIPAA safe harbor provisions do not set out to guarantee data protection like the newer modes of data protection; instead they only ensure a low risk of identification with the ultimate goal being "to balance the needs of the individual with the needs of the society." (Standards for Privacy of Individually Identifiable Health Information, 2000) The challenge is finding the "sweet spot" between protected data and useful data, while also understanding that this sweet spot changes for each dataset depending on what and how much information is available to the public. Furthermore, with rapidly evolving

technology, this sweet spot will continue to change over time. The amount of individual-level data collected by companies today is large and continuously growing. In fact, society may have already come to the point where the myth of the perfect population register is no longer a myth in the face of big data (Narayanan & Shmatikov, 2008).

While safe harbor continues to stand as the primary source of guidance for handling spatial health data, researchers continue to work with and against it in ways that reflect their understanding of the law and their data against a larger sociotechnical backdrop. As demonstrated by Malin et al (2011), there are ways to safely share more detailed data (i.e., age information) by coarsening the granularity of other data. From this example, we could assume that there are also ways to share finer-grained geographic data by censoring other elements in the data. Given that some pieces of information contribute more heavily to individual identification than others (i.e., DoB being more identifying than gender), we are left to ask some questions that, if answered, could help inform future approaches. Could a 5-digit ZIP code become innocuous without age information?[2] How many individuals can be uniquely identified by age and 5-digit ZIP code alone? What if all age and gender information were removed? Would a 5-digit ZIP code still have the power to identify an individual? In other words, is it reckless to share maps at the 5-digit ZIP code level if all other patient information is removed (i.e., only sharing 5-digit ZIP code and diagnosis)? What if these ZIP codes were aggregated together to form units that each contained 20,000 people within them? What would the risk for identification be? Of course, it is easier to ask these questions than answer them, but by examining the history of HIPAA and clarifying the important of 3-digit ZIP codes versus 5-digit ZIP codes, we have a stronger foundation for answering these questions. Until then, the safe harbor method stands as our primary mode of guidance and, two decades after its introduction, these guidelines do not meet the public's needs for data security nor researchers' need for useful data.

[2] HIPAA safe harbor requires that DoB be removed before data can be shared, but investigators are still allowed to share age information (as long as the person is under 89 years old).

6. Conclusion

Vague privacy provisions stand as an obstacle in the way of progress and pose a threat to public privacy by hindering the ways in which epidemiologists and geographers understand how to share spatial data. This paper promotes understanding of the HIPAA safe harbor provision by providing a comprehensive overview of the law while also presenting various expert perspectives and relevant studies that, taken together, show how alternative methods to safe harbor can offer researchers better data and better data protection. In particular, two different interpretations of the safe harbor rule exist—the 3-digit and the 5-digit zip code interpretation—and although 5-digit zip codes are not the intended level of aggregation under the rule, there is reason to believe that information can be safely shared in a map at this level. More research is needed in order to determine if the risk for individual identification is sufficiently low for maps shared at the 5-digit zip code level when DoB and gender are suppressed from a map's corresponding table. Much has changed in the twenty years since the introduction of the safe harbor provision, and yet it continues to be the primary source of guidance (and frustration) for researchers trying to share maps, leaving many waiting for these rules to be revised in accordance with the times.

Chapter 4. Regionalization with Self Organizing Maps for Sharing Higher Resolution Protected Health Information

Abstract

Background: This paper addresses the challenge of sharing finer-scale Protected Health Information (PHI) while maintaining patient privacy by using regionalization to create higher resolution HIPAA-compliant geographical aggregations. We use existing regionalization methods and introduce two novel regionalization approaches that integrate self-organizing maps (SOM) and then compare and contrast these methods in terms of their fitness for analysis and display. **Methods:** Four regionalization approaches based on different clustering methods (max-p-regions, REDCAP, and SOM variants of each) were used to each create a configuration of regions that aligns with census boundaries, optimizes intra-unit homogeneity, and maximizes the number of spatial units while meeting the minimum population threshold required for sharing PHI under HIPAA guidelines. The relative utility of each configuration was assessed according to: 1) model-fit characteristics using AICs and geosilhouettes and 2) region characteristics using compactness, homogeneity, and resolution. **Results:** Adding the SOM procedure to Max P resulted in statistically significant improvements for nearly all assessment measures whereas the addition of SOM to REDCAP primarily degraded these measures. The MSOM procedure's most notable improvements were seen in increases to average compactness and resolution. In contrast, RSOM produced degraded measures of average compactness, homogeneity, and resolution, only having a slight improvement in the variability of region size. The differences observed can be attributed to the different impacts of SOM on top-down and bottom-up regionalization procedures. **Conclusions:** Overall, REDCAP proves to be a superior approach to regionalization for the analysis and display of PHI, providing relatively high scores on characteristics most important for neighborhood health (compactness, homogeneity, and model-fit), as well as providing much finer regions than the standard approaches we rely on today. Additionally, MSOM—which provided the finest grained units—stands to offer an improved version of Max P for those who require a bottom-up procedure or can't access REDCAP.

1 Introduction

With the big data revolution underway, an inundation of data and new, powerful, computational tools have highlighted the need to find better ways to disseminate information about human health and well-being. This need was driven home by the Covid-19 pandemic and the desire on the part of many people and communities for fine-detailed information about local disease risk. One of the most common ways to share insights about human health is with data visualizations, including maps that are used to share geographically-linked, or spatial, information. The central challenge to mapping health data is the risk of disclosing highly confidential patient data by reporting the locations of people or cases in ways that make it easy to discover the identities and attributes of specific individuals. To overcome this challenge, researchers must strike a balance between sharing map data in a form that is useful—typically by offering finer-scaled data—and sharing map data in a form that protects patient privacy—typically by offering coarser-resolution data. The tension between needing fine and coarse data is a long-standing problem in health research. Geographic Information Science (GISc) approaches can help scholars solve this problem by offering innovative ways to share more useful data with research and policy communities while staying within the boundaries of privacy laws. In particular, regionalization — a geospatial approach of aggregating observations into new regions that can satisfy a variety of criteria—is a promising way to support better research with spatial health data. We examined several regionalization approaches for the case study of depression in the Twin Cities metro region of the United States

Before introducing these promising strategies, we must first understand the basic tenets of the privacy laws that regulate the way in which we share spatial data. The US Health Insurance Portability and Accountability Act (HIPAA) is one of the most commonly used set of guidelines for the use and sharing of spatial health data. In order to ensure that patient privacy is protected, the HIPAA privacy law provides rules to help data custodians understand precautions necessary to work with Protected Health Information (PHI) (HHS.gov, 1996). The main goal of HIPAA is to strike a balance between protecting the privacy of individuals and providing researchers with data that is still

useful. The most commonly used approach is termed the "safe harbor provision" of HIPAA, wherein individual locations must be aggregated to a polygon built from 3-digit zip codes that contains at least 20,000 people in addition to removing eighteen key identifiers, such as names, birthdates, and phone numbers. The idea here is that it is hard to identify specific individuals when there is not much known about them and they share the same characteristics with many other people. It is important to note that some confusion exists in regards to how geographic data should be aggregated to meet HIPAA standards (see Chapter 2 of this dissertation for more on this challenge).

In the US, health information is very often mapped at the county-level, which depending on context can be both too-restrictive and too-permissive from a HIPAA perspective. Many forms of health data are collected at the county level and much health policy and provision is a county responsibility. More broadly, people are arguably used to thinking about many issues in terms of county-by-county comparisons. Interestingly these county-level maps are not HIPAA compliant as in most instances a county shares more geographic information than what is allotted by the HIPAA safe harbor provision (which only permits data to be shared in the form of 3-digit zip codes and at aggregations of 20,000 persons or greater). And although many state agencies defer to state privacy laws as they are not regulated by HIPAA, if the state law is contrary to HIPAA, these agencies are required to follow the more stringent rule (which means that HIPAA could therefore restrict county-level data from being shared even by uncovered entities). Nonetheless, counties are one of the most commonly used units of display of health data in the US. The potential for improving public health and safety is often greater than the risk for individual identification (as with in an emergency outbreak or public health crisis), and therefore many agencies would conclude that the use of county-level data is justified given it serves communities while sufficiently protecting individual identities. While counties may not satisfy HIPAA because they are too-fine scaled, they are often also too coarse to be useful in many parts of the country. Consider metropolitan areas with large populations. Hennepin County Minnesota, for example, has a population of 1.3 million. This number is well over the 20,000 person threshold delimited under safe-harbor guidelines, so arguably it is possible for smaller geographies within Hennepin

County to be safely shared without putting individuals at risk for identification (as long as these smaller geographies still had more than 20,000 people within them).

Here we explore the potential for a less stringent but arguably still-valid interpretation of the safe-harbor provisions. We use *regionalization*, or zone design, as a way to build better HIPAA-compliant maps. Regionalization is a way of developing spatial units that satisfy key elements of HIPAA while also being more useful for research and policy. We explore a generic approach that allows for the aggregation of health records into any geocode (meant in the sense of any arbitrary region that encompasses at least 20,000 people) rather than only 3-digit ZIPs (Browne et al., 2014; Jung & El Emam, 2014; Mu et al., 2015). The reasons for exploring alternates to the 3-digit ZIP code are twofold: 1) it allows for finer resolution data therefore permitting the exploration of more computationally intensive techniques, and 2) it is more practical in the sense that very few people share data at the 3-digit ZIP code level because this type of geography is too coarse-grained (i.e., it often covers large expanses) and is unfamiliar to most people. Regionalization offers units that align with census boundaries, optimize intra-unit homogeneity, and maximize the number of spatial units while meeting the minimum population threshold required for sharing PHI under HIPAA guidelines.

Regionalization is a geospatial analytical process that builds custom regions from underlying data to suit a specific function or for the display of specific data. This approach gives researchers control over the shape, size, and demographic makeup of the resultant regions. Regionalization can aggregate underlying units (or observations groups into spatial units) while optimizing an objective function based on the investigator's research needs (Openshaw, 1977). A common example is aggregating units of relatively fine-scaled census geography, like blocks or block-groups, into a set of larger regions for analytical purposes. One important form of optimization is developing regions that maximize homogeneity within regions and maximize heterogeneity between them in order to support other forms of analysis. For example, regionalization can increase the power of statistical analyses by developing better observations. Data aggregation reduces the margins of error from insufficient samples (which can be quite large especially within

fine-scale units such as census block groups) by ensuring that units minimize artificial heterogeneity (Folch & Spielman, 2015). It may also offer to strengthen other kinds of analysis by making regions more homogenous *among* one another (as opposed to within each unit) as a way of controlling a given variable. For examples, by building regions that are uniform in terms of median household income we can, in essence, help control for potential confounding when mapping related variables like green space or pollution (Krzyzanowski et al., 2019).

Regionalization holds many potential advantages for the analysis and sharing of PHI under the aegis of HIPAA's safe harbor provision. In spite of the general importance of regionalization to spatial analysis, its use in the context of privacy protection has been very limited. Very few studies have explored the use of regionalization as a means to create units that meet data privacy regulations (Croft, 2016; Mu et al., 2015; Wang, Guo, & McLafferty, 2012). These studies used regionalization (Wang et al., 2012) or multiple regionalization strategies (Croft, 2016; Mu et al., 2015) to create configurations that reduce the amount of suppression required, maximize the number of regions, or maximize the compactness of regions. Despite many insights, at the time of writing, none of these studies offer publically available tools or workflows that can be easily used by others. Furthermore, these regionalization procedures were designed for specific use cases with the goal of restricting the extent to which the geographic aggregations could be refined either to achieve sufficient anonymity of microdata (Croft, 2016) or reliable risk estimates for low-incidence disease such as cancer (Mu et al., 2015; Wang et al., 2012). Therefore, if these regionalization procedures were eventually shared with the public, the scripts and workflows would still require some sort of modification in order to achieve the finest-grained map units possible for sharing PHI.

The present project uses regionalization to create higher-resolution HIPAA-compliant regions in order to address the need to share finer-scale maps while adhering to privacy standards. This research offers several significant advances in the use of confidential health data. First, it addresses a real need for a greater variety of ways to work with, present, and understand PHI. Second, it advances knowledge of how we could use

regionalization to analyze and report PHI in ways that satisfy HIPAA guidelines, or more generally, any rules that specify population thresholds and geographical limits. By extension, this work points the way towards sharing data with the community at a meaningful resolution without breaching HIPAA privacy regulations. Third, our case study uses a real public health dataset (depression diagnoses) to assess best-fit among different regionalization outputs. Therefore, in addition to advancing the theory and method of data-sharing and visualization, there is potential to provide innovative tools to facilitate dissemination of fine-scale information within patterns of depression to the community.

2 Methods

We test several different regionalization methods in terms of their potential to develop HIPAA-compliant maps in the context of neighborhood-level depression risk in the Twin Cities region of Minnesota, United States. The regionalization procedures and assessments require the use of two different data sets, patient-level data on depression and socioeconomic data from the US Census. We tested two basic forms of regionalization, heuristic (Max P) and hierarchical (REDCAP), and integrated self-organizing maps with each, so we have four basic modalities. We assessed each of these four types of regionalization with two measures of model-fit (Akaike Information Criterion and Geosilhouettes) as well as three measures of region characteristics (homogeneity, resolution, and compactness).

2.1 Data and the Twin Cities region

We conducted our analyses for a case study set in the Twin Cities of Minnesota. The Twin Cities metropolitan region includes Minneapolis and St. Paul and encompasses the seven counties that contain and surround these cities (Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington). The seven-county region is commonly used in research as it provides a broad range of human and environment characteristics from which to explore.

We use two basic datasets: socioeconomic data from the US Census and patient-level data from a Twin Cities' health system. We used the socioeconomic data to guide the regionalization, especially to achieve homogeneity and a minimum population, and to create a simple model of depression for assessing model-fit. We used 2010 census data, including median household income and education as measured as number of persons per census block group with a bachelor's degree or higher. Median household income and educational attainment together have been shown to be an adequate measure of SES (Gerber et al., 2008; Roblin, 2013; Siahpush, Heller, & Singh, 2005). In addition to these census data, the project used the Fairview Health system's data from the Academic Health Center Clinical Data Repository to assess model-fit in each of the configurations (Fairview Health Clinical Data Repository, 2019). Patient data included electronic health records of 97,432 outpatient visits between 2010 and 2018. Patients were included if they were over the age of 18 and had at least 1 depression diagnostic code (ICD-9 code: 296.20, 296.22, 296.23, 296.30, 296.32, 296.33, 311; ICD-10 code: F32.XX, F33XX). For patients who presented more than one diagnostic visit between 2010 and 2018 only the first instance was kept. The prevalence of depression was calculated as the number of cases divided by the total population within each unit.

2.2 Regionalization overview

There are many different regionalization methods. The present study required a regionalization method that: 1) guaranteed contiguous regions (so as to not have gaps and to have neighborhoods), 2) constrained regions to align with census boundaries (because we use census data at various scales), 3) incorporated a population minimum (for privacy protection), and 4) did not limit the number of regions (in order to create as many as possible). We examined methods that meet these requirements and exemplify the three main kinds of generic clustering approaches (in the sense that regionalization is a specialized instance of the more general set of clustering approaches used across the sciences), namely heuristics, hierarchical clustering, and neural networks (Dao & Thill, 2018). For heuristics and hierarchical clustering respectively, we chose two widely accepted current methods, Max P Regions and REDCAP, while for neural networks we

turn to self-organizing maps (SOM). SOM must be tied to another regionalization method (it is a clustering approach but not a spatial regionalization method) and so we introduce two novel methods—variants of Max P and REDCAP that integrate SOM. We refer to these new regionalization approaches as MSOM (Max P + SOM) and RSOM (REDCAP + SOM). In sum, we have four key approaches: Max P, REDCAP, MSOM, and RSOM.

Max P Regions is so named because it solves the 'p-regions problem,' which attempts to find the maximum number of contiguous areas that can be created from a study area while optimizing an objective function. The method begins with an initial random seed point, merging contiguous base units until a satisfactory solution is achieved or—if the configuration is determined to be unfeasible—a new random seed point is chosen and the process starts over again. Max P can be thought of as a bottom-up procedure as it starts by iteratively linking together smaller units into larger regions and continuing to build-up until the final solution is reached. Because Max-p partitions zones by the means of linear integer programming (LIP) which can be computationally intensive, it is vulnerable to becoming trapped in a local minimum (meaning that it could converge on a less than optimal configuration) (Li, Church, & Goodchild, 2014; She, Duque, & Ye, 2017). In order to address this, there are many heuristic approaches that can be added to Max P Regions including simulated annealing and tabu search—both of which prolong the search for an optimal solution by allowing for non-improving moves.

REDCAP stands for Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning and describes a family of six methods that partition according to different attribute homogeneity (constraining) and contiguity rules (Guo, 2008). REDCAP is a hierarchical procedure, meaning that sub-regions are nested within larger areas. Unlike Max P, REDCAP can be thought of as a top-down procedure that starts by creating a minimum spanning tree (MST) by growing branches, or links, between contiguous units that are the most similar to each other in attribute space (i.e., linking block groups that are most similar in median household income). Once all units are connected in the tree, the branches are progressively cut with the aim to maintain the smallest overall sum of squared deviations possible. In other words, the first cut would be

between the two connected units that, when separated, produce two subtrees, or nests, that have smaller within group variances than the initial tree had prior to the cut. Culling of the tree continues until subsequent cuts can no longer result in an improvement to the final solution and the units are merged within their respective subtrees, becoming a new region.

Self-Organizing Maps (SOMs) are a type of artificial neural network that is used to transpose complex high-dimensional data (including multivariate components across space and time) into a one or two dimensional map-like surface that highlights the strongest aspects of the dataset in different places in space and/or time. The neural network strategy places nodes (neurons) across a data set, and then uses those nodes to build a network of relation by assigning weights to each node based on its similarity to a vector, or the starting node. The vector in SOM is chosen at random at the start of SOM training. The SOM algorithm is adaptive in the sense that it relies on unsupervised learning, specifically competitive learning, to create the neural network that finds the best matching unit (or the node that is most similar to the vector) and then transposes the winning characteristics onto a map-like surface. The Geo-SOM method, developed by Bação et al, accounts for space by incorporating a geographical tolerance or k parameter which restricts the search for the winning node to a neuron's geographical neighbors (2004). The Geo-SOM method has shown potential to be integrated into regionalization processes as it creates delineations between homogenous areas (Relia, Akbari, Duncan, & Chunara, 2018; Bação et al, 2005). In particular, we turn to guidance from Bação's et al's conference paper which suggested that ridges between places where high values meet low can be easily distinguished after transposing the u-matrix onto a geographical surface, and that these areas of change can be used to delineate homogenous areas (2005). With this knowledge, we introduce our two novel methods for partitioning areas by integrating GeoSOM with Max P and REDCAP regionalization to create the MSOM and RSOM methods.

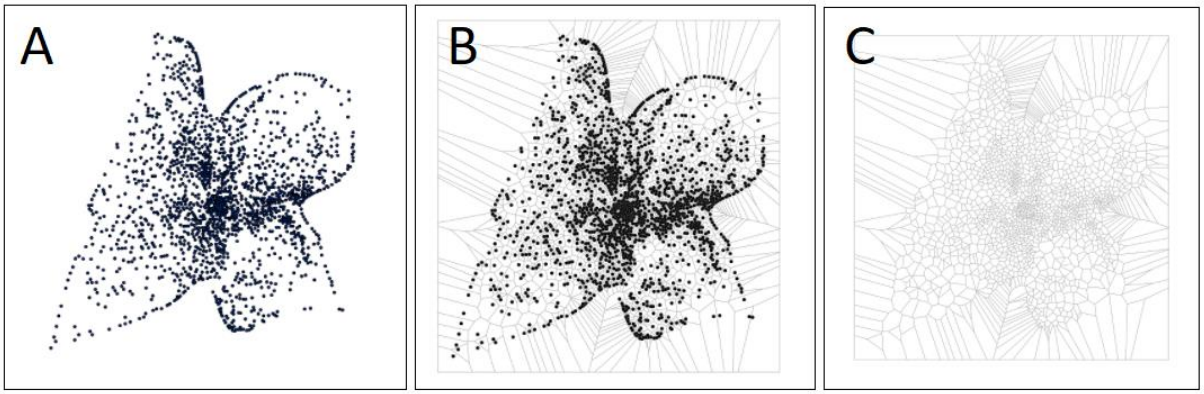
2.3 Regionalization specifics

The base units, or building blocks, for all four regionalization approaches were set to block groups. We repeated all regionalization processes thirty times to create thirty different configurations for the assessment stage of the analysis.

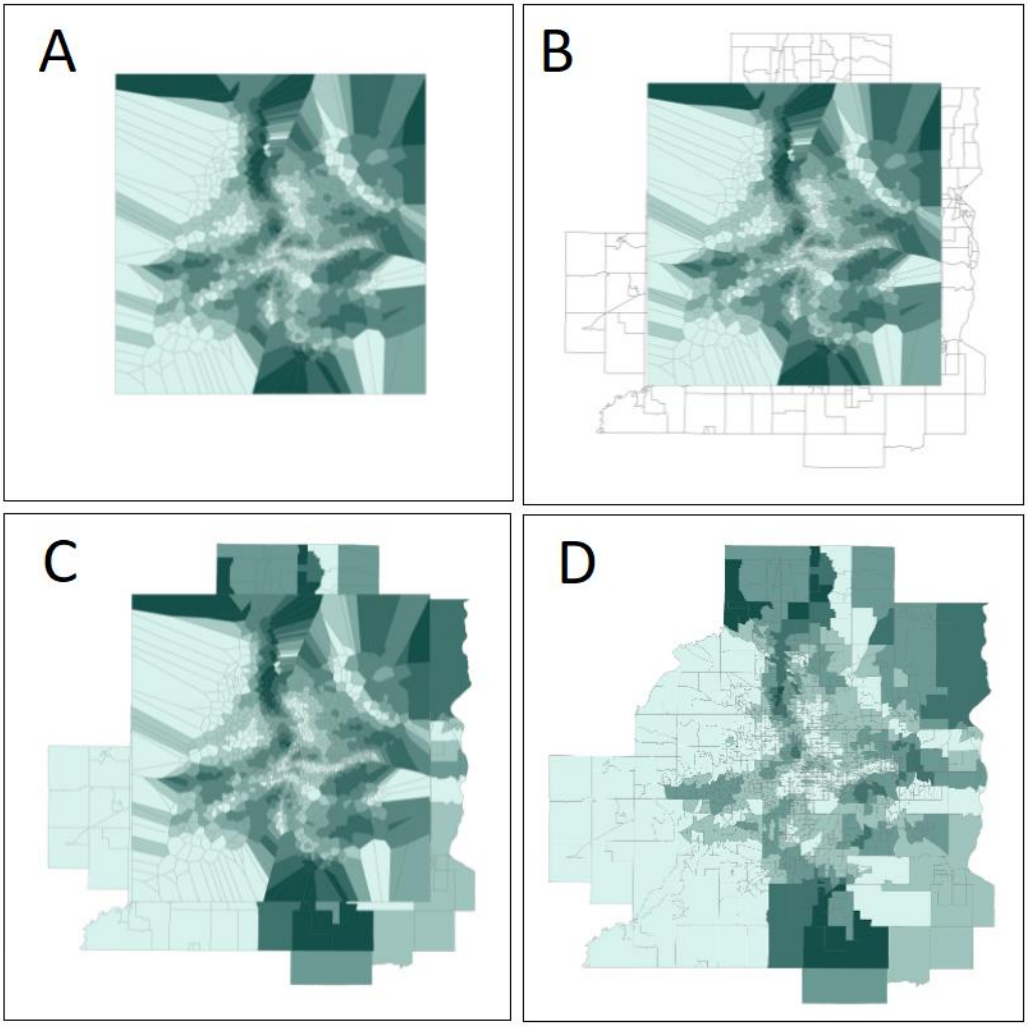
Max P Regions. The Python library *ClusterPy* (Duque, Dev, Betancourt, & Franco, 2011) was used to implement the max-p-regions regionalization. The regionalization was used to optimize areas to be homogeneous according to median household income and educational attainment. A floor constraint of 20,001 was used to ensure that every region contains a population of more than 20,000. The maximum number of iterations was set to 100 in order to increase the likelihood of achieving the maximum number of regions. The tabuLength parameter was set to the default value of 85—limiting the number of non-improving moves.

REDCAP. Guo's REDCAP software was used to implement the REDCAP method (2008). The regionalization method was set to average-linkage clustering with full-order constraining. These parameter settings were selected in accordance with previous work that suggests that full order constraining with average-linkage clustering maximizes the number of regions produced (Kugler, Manson, & Donato, 2017). The weights matrix used in the regionalization was built using rook contiguity and the regionalization process was set to optimize by the income and education variables (which were given equal weight). In order to ensure that the output meets HIPAA safe harbor standards, regions were set to contain a minimum of 20,001 persons. Unlike Max P Regions, REDCAP creates the exact same output with each run. Therefore, in order to produce 30 slightly different configurations for our assessment, the smoothing setting needed to be changed after each execution. Adaptive kernel smoothing settings of 1 through to 28 were used, in addition to using no smoother and having one configuration that relied on empirical Bayesian smoothing (set to have three neighbors). We exhausted all of the available smoothing settings within REDCAP in producing the 30 different configurations.

MSOM. Our MSOM procedure involves integrating Bação's et al's Geo-SOM method with Max P regionalization. The initial steps require the Geo-SOM procedure to build a u-matrix which is transposed onto the geographical surface. This was achieved within Matlab (Bação's et al's Matlab routines are available at <https://www.novaims.unl.pt/labnt/geosom/index.htm>) and R Project (our script available at <https://z.umn.edu/SOMregionalization>). The Geo-SOM was trained with a map initialization of 100 x 100; the geographic radius was set to 2 neighbors and a sheet hexagon lattice was used. Latitude and longitude coordinates defined the geographical components used in training the GeoSOM, while median household income and educational attainment served as the non-geographical components. GeoSOM was iterated 30 times (to obtain 30 different u-matrices). Each of the thirty u-matrix tables were joined with their corresponding latitude and longitude component table to create 30 complete output tables. Figure 4.1A illustrates the spatial spread of the nodes for one of the thirty GeoSOM executions which can be observed from plotting the X and Y coordinates. This process was streamlined via an R script that preprocesses GeoSOM output to prepare it for regionalization by joining the tables, formatting variable headings, removing records with NaN values, plotting the nodes, and building Thiessen polygons around each node (script available at <https://z.umn.edu/SOMregionalization>). Within this script, Thiessen polygons are assigned the u-matrix value of the node that falls within it (Figure 4.2A) and then projected onto a map of the Minneapolis metro area. U-matrix values were then transposed onto the map surface by assigning the u-matrix value of the Thiessen polygon to the block group whose centroid falls within it (Figure 4.2B-D). This block group-level map was then used as input for the Max P Regions regionalization procedure. By setting the regionalization to optimize according to u-matrix values, we partitioned the study area according to the homogenous areas identified by Geo-SOM while ensuring a minimum population of 20,001 persons per unit. Supplemental illustrations and animations detailing this process can be found at (<https://z.umn.edu/SOMregionalization>).



4. 1 A) Nodes for one of the thirty GeoSOM executions. B-C) Thiessen polygons built around each node.



4. 2 A) Thiessen polygons each assigned the u-mat value of the node that falls within it. B) Thiessen polygons projected onto a map of the Minneapolis metro area. C) U-matrix values transposed onto the map surface by assigning the u-matrix value of the Thiessen polygon to the block group whose centroid falls within it. D) Final u-matrix map at the block group level.

RSOM. 30 separate u-matrices were transposed onto a geographical surface in the same manner as described in for MSOM. Then these 30 u-matrix maps were used as input for the REDCAP procedure. The RSOM procedure relied on the same parameter settings that were used within the REDCAP procedure (see REDCAP section), with the only difference being that the regionalization process was set to optimize by u-matrix values instead of the income and educational attainment variables.

2.4 Assessment procedures

Assessment is a key step in evaluating the utility of configurations created from different regionalization approaches with respect to their fitness for use in developing analytical frames while protecting patient confidentiality. In simple terms, we adopt a two-pronged assessment of each regionalization (Table 1), focusing on model-fit characteristics and then on spatial measures.

Model-fit Characteristics	**AIC**
	Geosilhouettes
Spatial Measures	**Compactness**
	Homogeneity
	Resolution

Table 4.1 Assessment approaches.

First, we determine how well a given configuration supports a simple model of health with a standard model-fit measure, the Akaike Information Criterion (AIC), and then the more spatially-oriented geosilhouette approach. The AIC is traditionally used as a means of model selection, or finding the most parsimonious model from a set of candidate models that use different covariates and/or interaction terms, but it can also be used to assess relative goodness-of-fit in a set of models that use the same variables but different geographic scales (Cabrera-Barona et al., 2016). We used multilevel modeling techniques to develop a simple model of linear regression of greenspace and air quality on risk of depression and use this model to compare the goodness-of-fit between the four

different regionalization methods. The geosilhouette approach did not rely on this model and is an assessment of depression risk and space alone.

AIC. The Akaike Information Criterion (AIC) is a well-established and widely supported in the modeling literature (Burnham & Anderson, 2004; Cabrera-Barona, Wei, & Hagenlocher, 2016; Rose & Nagle, 2017). In order to determine which configuration provided the best model-fit, we used the four regionalizations to regress depression risk on greenspace and then compared the AICs across each model. Lower AIC values indicate more parsimonious models or a relative better fit that strikes a balance between over and under-fitting the data. In order to do this, we first used multilevel regression modeling to select a suitable model from a set of possible models describing the relationship between depression risk, greenspace, air quality, and median household income. Multilevel modeling revealed a more parsimonious model could be obtained by adding air quality as a covariate. However, the addition of median household income to the model did not improve the model-fit. For this reason, the final model maintained air quality but not income ($Y_{risk} = a + bx_{greenspace} + bx_{air\ quality}$). After selecting the model, we performed 30 regression analyses for each of the 4 regionalization approaches. The average AIC was compared across each regionalization approach to assess relative model-fit and determine which of the four configurations is (on average) better suited for modeling the relationship between depression and greenspace (as determined by the model with the lowest average AIC). Regression modeling is commonly used in geographic analyses of aggregated data (Fei et al., 2016; Hallowell, Robb, & Kintziger, 2018; Iroh Tam, Krzyzanowski, Oakes, Kne, & Manson, 2017).

Geosilhouette. Geosilhouettes are a geographic approach to model-fit specifically created for model-fit for geographical units assigned to larger clusters (Wolf, Knaap, and Rey, 2021), which in our case means regions built from regionalization. Traditional measures of goodness-of-fit largely focus on attribute homogeneity and ignore or simplify the role of space. For example, Rousseeuw's 1987 original silhouette model measures a single observation's goodness-of-fit to its region (relative to another region) using Euclidean distances. In contrast, Geosilhouettes uses a modified definition of distance and similarity

in order to incorporate a more meaningful measure of joint geographic-attribute similarity. There are two kinds of geosilhouette's: path and boundary silhouettes. In this study, we use the path geosilhouette model, which uses the path dissimilarity distances to assess how well a regionalization strategy places a block group into a region, given that the block group's next-best-connected (NBC) region may be further away. The path dissimilarity model accounts for the dissimilarity between the block group and the block groups within its NBC region by looking at the total attribute dissimilarity along the path that connects them. This is how the path silhouette modifies the distance metric of the original silhouette to account for geography. It is a silhouette score that uses the length of the dissimilarity paths within the computation. A path silhouette score is calculated for each block group taking into account the joint spatial–social similarity of the block group to the block groups contained within its NBC region. When the silhouette score is close to 1, that means that the block group has a short attribute-weighted path to its NBC region. In other words, the block group is physically close to its NBC region and/or very similar in attribute value to it. If the silhouette score is close to -1 that indicates that the block group is better connected with block groups in its own region compared to the block groups within its NBC region. In this study, path geosilhouettes scores were calculated for each block group and then the average geosilhouette score was calculated for each region for all 30 runs. All calculations were executed using the PySAL library that is freely available to the public from https://pysal.org/esda/notebooks/geosilhouettes.html.

The second set of approaches for assessing regionalization involves examining how well each method can produce regions that are meaningful or useful according spatial measures. Model-fit characteristics help determine the extent to which a regionalization help capture an important modeled relationship, but the best-fitting configuration for one model is not necessarily the one that will capture a wider array of potential relationships (Openshaw, 1977). Therefore, it is helpful to examine each regionalization according to how well it offers more generic desirable neighborhood characteristics. We assess regions with three of the most commonly accepted and longstanding measures in the region and cluster assessment literature, namely resolution (shape and size), intra-unit homogeneity,

and compactness (Openshaw, 1977; Openshaw, 1983). Note that we do not explicitly assess the extent to which method enhances data privacy because it is held constant across methods by being integrated into each regionalization process itself via the population minimum threshold of 20,000 people.

Homogeneity. Pinzari and others (2018) developed a measure appropriate for assessing homogeneity of regions built from aggregation. Their homogeneity index accounts for how an attribute is distributed across ordinal categories (in deciles) within each region rather than only accounting for the range of the data (for a thorough explanation, see Pinzari et al. (2018)). Homogeneity indices were calculated for each region and then averaged across the entire configuration for all 30 runs. All calculations were executed using an R script that is freely available from https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-018-0162-8#Sec16.

Resolution. Overall resolution for each configuration was determined according to the size of regions (measured as polygon areas) and the variance of the region areas within each configuration. In terms of spatial modeling, configurations with a smaller average area value are generally considered more desirable as smaller units are generally associated with higher resolution and offer greater spatial specificity in terms of measuring features on the ground (Goodchild, 2011). Similarly, configurations with a smaller variance of region areas are considered more desirable for spatial modeling because there is a more consistent region size across the realization. Variance provides a simple measure of the extent to which the level of detail in one part of the configuration matches all other parts of the configuration.

Compactness. Average compactness of each configuration was assessed with the isoperimetric ratio, perhaps the most widely-accepted standard for compactness (Li et al., 2013; Osserman, 1978). This compactness measure is calculated by dividing the area of a region by the area of a circle with the same perimeter as the region (Kugler et al., 2017). Thus, a high isoperimetric ratio indicates more compact regions and is generally

seen as more desirable in how it avoid sprawling shapes. The overall compactness of the configurations was calculated as the average of compactness scores over all 30 realizations.

3 Results

We assessed the four regionalization modalities (Max P, REDCAP, MSOM, and RSOM) across the five assessment measures, namely model-fit characteristics (AIC and geosilhouettes) and region characteristics (compactness, homogeneity, and resolution). We use a mix of statistical and graphical reporting methods, supplementing use of test statics and p-values with group means, effect sizes, maps, and graphics. Our primary statistical approach was F-tests with post hoc comparisons of the four regionalization approaches. Levene's F test revealed that the homogeneity of variance assumption was not met for any of the assessment measures ($p < 0.05$). Since the assumption of homogeneity of variance was not met for this data, we used Welch's ANOVA (Welch's F test) followed by Games-Howell post hoc comparisons. Furthermore, normality testing revealed that the homoscedasticity assumption was met for all measures except for the geosilhouettes. For this reason, geosilhouette scores were analyzed with nonparametric tests, the Kruskal-Wallis test and Dunn's test with a Bonferroni correction.

There are many potential comparisons among the four methods but since the four are not independent (given how SOM modifies other methods), we focus on comparing the two parent approaches (Max P and REDCAP) and the parent and offspring approaches (Max P vs MSOM and REDCAP vs RSOM). Even though we focus on three comparisons, we opted for a conservative alpha value for our post hoc tests (setting alpha to the Bonferroni adjusted .008 for six comparisons (.05/6)). We do this because MSOM and RSOM stand best to be compared to their parent regionalization approaches; however, we acknowledge that others may have interest in comparisons that we did not find that interesting. Therefore, results from all 6 comparisons can be found in figure 5.1 in the appendix.

Table 4.2 The effect of regionalization on region characteristics.

Groups	Compactness	Homogeneity	Region Size (km^2)	SD of Region Size (km^2)	AIC	Geosilhouette
Max P	.20 ± .001	.70 ± .003	28.1 ± .23	117 ± 1.2	-598.5 ± .98	-.136 ± .001
MSOM	.23 ± .001***	.63 ± .002***	26.6 ± .19***	102 ± 1.2***	-626.9 ± 1.6***	-.139 ± .001
REDCAP	.27 ± .004***	.64 ± .002***	30.0 ± .34***	123 ± 1.2*	-554.5 ± 2.4***	.013 ± .002***
RSOM	.25 ± .003###	.62 ± .002###	29.4 ± .33	114 ± 1.3##	-558.5 ± 2.3	.000 ± .002

Data were presented as means ± standard error of the mean. All data, except for the geosilhouette measure, were analyzed with Welch's ANOVA followed by Games-Howell post hoc multiple comparison test. Geosilhouette data was analyzed with the Kruskal-Wallis test followed by Dunn's test for post hoc comparisons with a Bonferroni adjustment.

***,### Significant difference values $p<.0001$

**,## $p<.005$

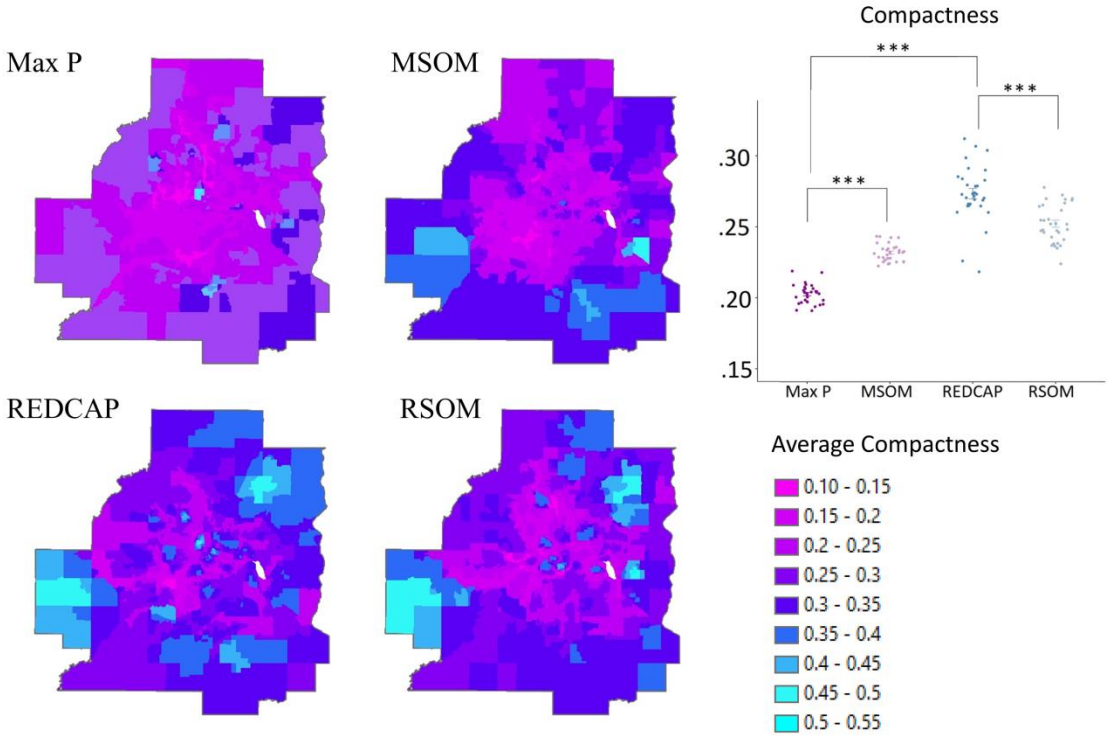
# $p<.008$ (Bonferroni correction for six comparisons)

*A significant difference relative to Max P.

A significant difference relative to REDCAP.

3.1 Spatial measures: Compactness

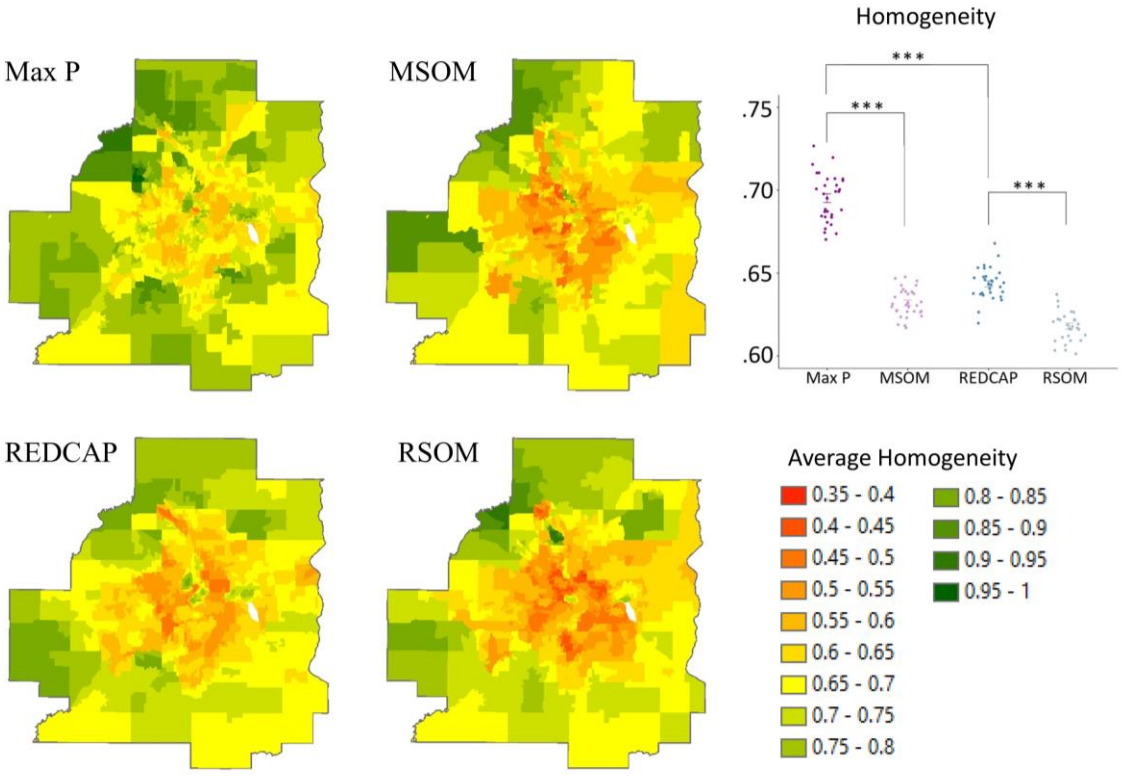
Welch's ANOVA determined that there was a statistically significant difference in mean compactness scores between the regionalization methods Welch's $F_{(3, 61.14)} = 203.25$, $p < .0001$, $\omega^2 = .90$, 90% CI [0.87, 0.93]. Subsequent post hoc comparisons with Games-Howell revealed that REDCAP produced configurations with significantly higher mean compactness scores (.2732 ± .0037) compared to all other regionalization methods including Max P (.2026 ± .0013). In terms of the differences observed between parent (without SOM) and offspring regionalization (after SOM integration), we found that adding SOM to the Max P procedure (i.e., MSOM) resulted in a 14% increase in mean compactness to .2313 ± .0012 ($p < .0001$). In contrast, adding SOM to the REDCAP procedure (RSOM) resulted in an 8% reduction in mean compactness scores (.2521 ± .0026) which was statistically significant ($p < .0001$).



4. 3 150-meter resolution raster maps of the average cell value calculated from stacking 30 rasterized compactness score maps for each of the four regionalization strategies (Max P, MSOM, REDCAP, and RSOM). A single point on the scatter plot represents the average compactness score of one map configuration. There are 30 points in each regionalization group.

3.2 Spatial measures: Homogeneity

Results from Welch's F test determined that there was a statistically significant difference in mean homogeneity index between the regionalization methods Welch's $F_{(3, 63.52)} = 191.84$, $p < .0001$, $\omega^2 = .89$, 90% CI [0.85, 0.92]. Post hoc comparisons (with Games-Howell) revealed that Max P produced configurations with significantly higher mean homogeneity indices (.6952 ± .0027) than all other regionalization methods including REDCAP (.6435 ± .0018). Additionally, adding SOM to the Max P procedure (MSOM) resulted in a 9% decrease in average homogeneity (.6321 ± .0016) which was statistically significant ($p < .0001$). The same was found when adding SOM to the REDCAP procedure (RSOM), which reduced average homogeneity to .6165 ± .002 ($p < .0001$).

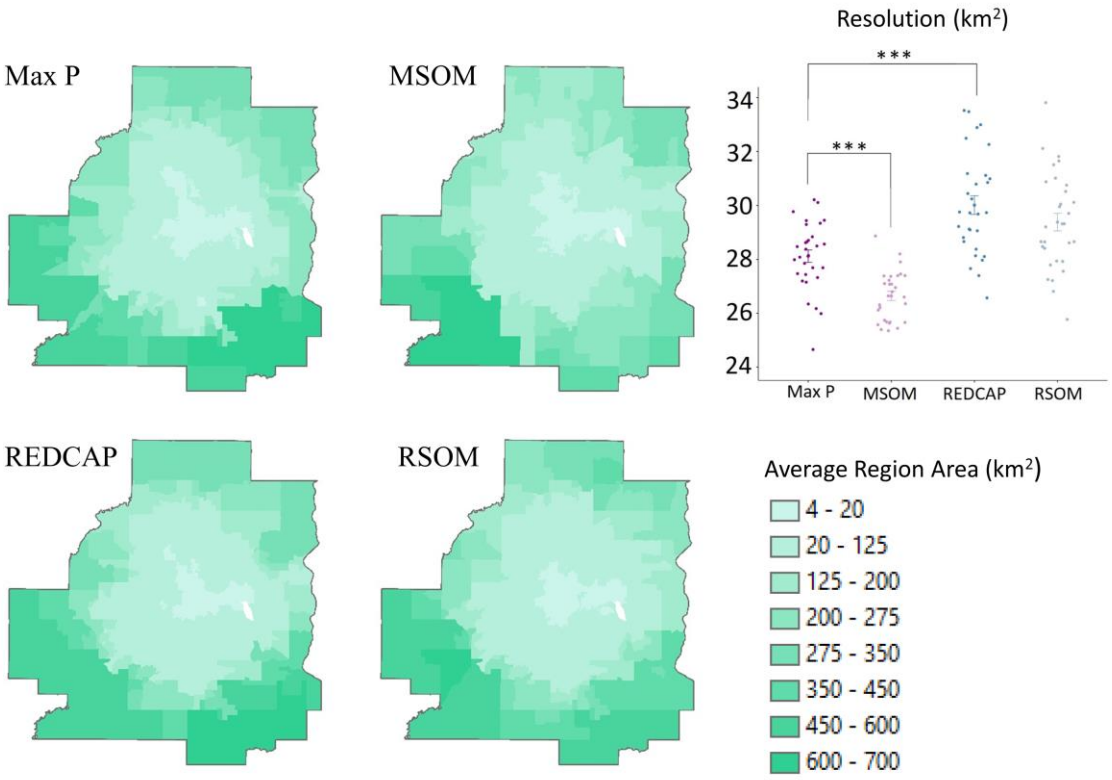


4. 4 150-meter resolution raster maps of the average cell value calculated from stacking 30 rasterized homogeneity indices maps for each of the four regionalization strategies (Max P, MSOM, REDCAP, and RSOM). A single point on the scatter plot represents the average homogeneity of one map configuration. There are 30 points in each regionalization category.

3.3 Spatial measures: Resolution

Welch's ANOVA results determined that mean resolution (according to average area of the regions within a configuration) differed among the regionalization methods, in that Welch's $F_{(3, 62.65)} = 36.0$, $p < .0001$, $\omega^2 = 0.61$, 90% CI [0.48, 0.70]. Games-Howell post hoc comparisons revealed that the Max P procedure produced configurations with significantly finer-grainer units, or units with lower average areas, ($28.1 \pm .2316$ km^2) than REDCAP ($30.0 \pm .3439$ km^2). Additionally, adding SOM to the Max P procedure (MSOM) reduced the average area even further to $26.6 \pm .172$ km^2 ($p < .0001$). No statistically significant difference in resolution was observed between REDCAP and RSOM ($29.4 \pm .3308$ km^2, ($p = .55$). The average number of regions per configuration corroborates these findings. Whereby, MSOM produced the most regions (117.066), followed by Max P (112.633), REDCAP (103.8), and RSOM (99).

ANOVA results also showed that the mean variability of regions areas (standard deviation of region areas) differed among the regionalization methods Welch's $F_{(3, 64.23)} = 54.14$, $p < .0001$, $\omega^2 = 0.70$, 90% CI [0.59, 0.77]. Games-Howell post hoc comparisons revealed that the Max P procedure produced configurations with statistically significantly lower average variation in region size (117 ± 1.23 km^2) than REDCAP (123 ± 1.37 km^2; $p = .009$). And additionally, adding SOM to the Max P procedure (MSOM) reduced the average variability in region area even further to 102 ± 1.09 km^2 ($p < .0001$). Furthermore, although no difference was found in average region size between REDCAP and RSOM (paragraph above), a statistically significant difference in average variability in region size *was* observed between these two approaches (116 ± 1.29 km^2, ($p = .004$)). See figure 5.2 in the appendix for a scatter plot for the standard deviation of areas.



4. 5 150-meter resolution raster maps of the average cell value calculated from stacking 30 rasterized average area maps for each of the four regionalization strategies (Max P, MSOM, REDCAP, and RSOM). A single point on the scatter plot represents the average area of all of the regions contained within one map configuration.

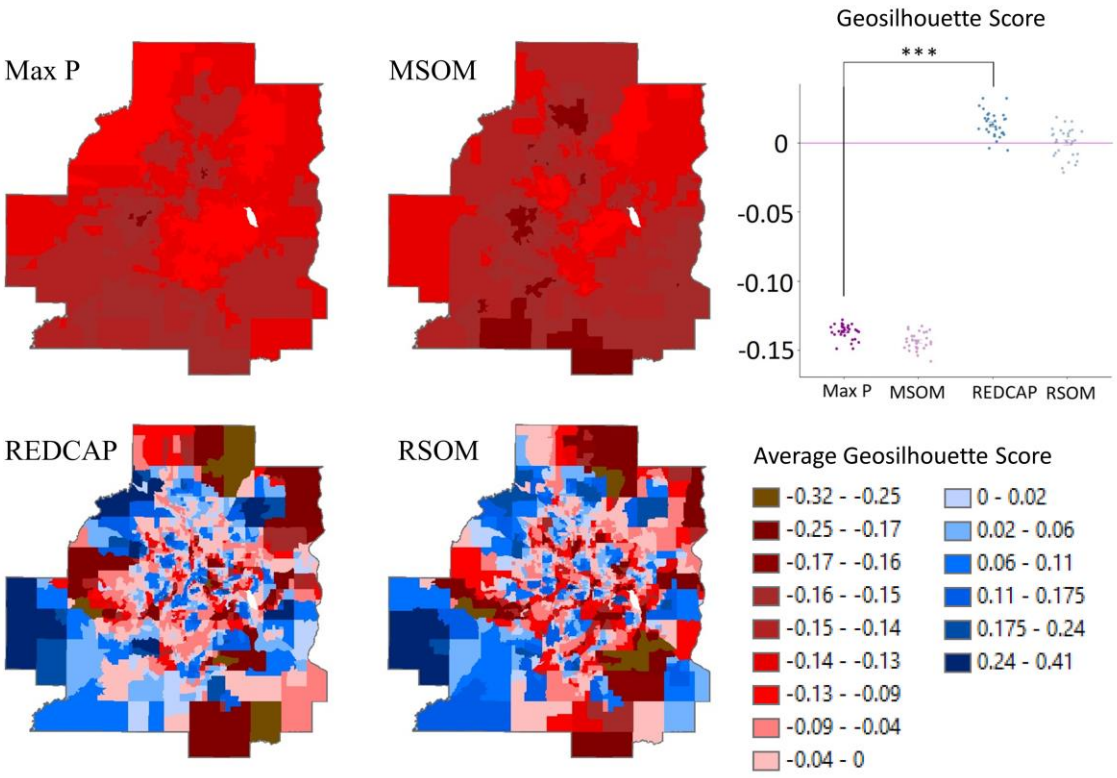
3.4 Model fit: Akaike Information Criterion.

Welch's F test results determined that mean AIC (outputted from a linear regression of greenspace and air quality on risk of depression) differed among the regionalization methods Welch's $F_{(3, 57.51)} = 196.4$, $p < .0001$, $\omega^2 = .91$, 90% CI [0.87, 0.93]. Lower AIC values indicate more parsimonious models and therefore lower AIC values are more desirable (or higher negative values). Post hoc comparisons revealed a statistically significant difference between Max P (-642.1 ± 14.5) and REDCAP (-634.6 ± 12.1). Furthermore, there was a significant difference in average AIC after adding SOM to the Max P procedure (MSOM) -653.8 ± 10.5. However, adding SOM to REDCAP did not have a statistically significant effect on the average AIC (-547.5 ± 6.6, p=0.3).

Recall that prior to AIC comparison, multilevel modeling was used to select which covariates should be included in our model of depression risk. If, however, someone were to compare AICs across regionalization types using a model with different covariates and/or interactive terms, the relative pattern observed across regionalization types may change. In fact, we found a very different pattern when comparing the AICs in a simple bivariate model of depression risk and greenspace Welch's $F_{(3, 61.49)} = 34.3$, $p < .0001$, $\omega^2 = .60$, 90% CI [0.47, 0.69]. These results are important as they demonstrate that using AIC as a metric for comparing relative model-fit between regionalization strategies could be considered somewhat questionable because the AIC is rather sensitive to the model used. For this reason, we recommend that the AIC be only used to assess relative model-fit in geographic regionalization studies when the model is pre-specified and investigators are confident in their choice of variables. See figure 5.2 in the appendix for a scatter plot for the average AICs.

3.5 Model fit: Geosilhouettes

Kruskal-Wallis test results determined that mean geosilhouette scores (derived from modeling depression risk) differed among the regionalization methods Kruskal-Wallis $\chi 2 = 97.24$, $p < .0001$, $df=3$, $\varepsilon^2 = .82$, 90% CI [0.79, 0.85]. Bonferroni adjusted post hoc comparisons with Dunn's test revealed that REDCAP had a higher mean geosilhouette score (better model-fit) than Max P (.013 ± .001 vs -.136 ± .001, respectively ($p <.0001$)). No significant difference was observed between Max-P and MSOM (MSOM= -.139 ± .004, $p = 0.2$) or between REDCAP and RSOM (RSOM=.004 ± .002,$p=.09$). When using median household income instead of depression risk as the variable of interest in the geosillhouettes model, the relative pattern observed between regionalization approaches remained the same.

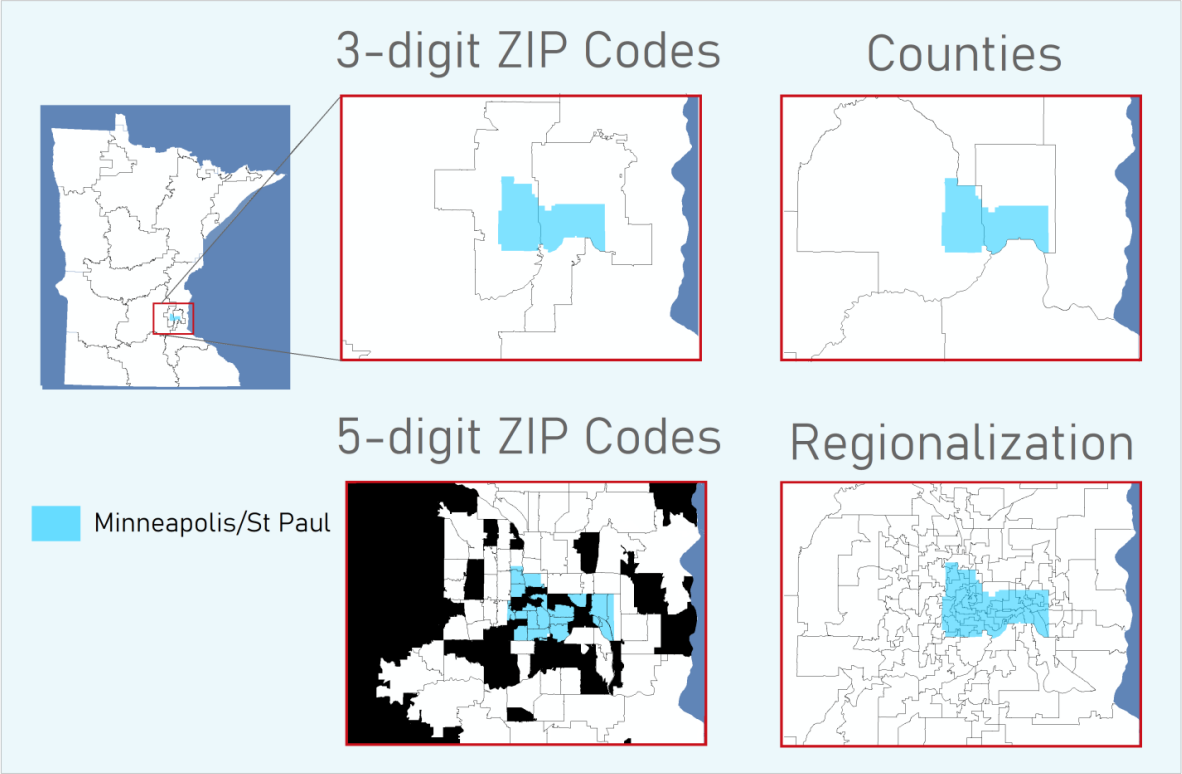


4. 6 150-meter resolution raster maps of the average cell value calculated from stacking 30 rasterized geosilhouette score maps for each of the four regionalization strategies (Max P, MSOM, REDCAP, and RSOM). A single point on the scatter plot represents the average geosilhouette score of one map configuration. There are 30 points in each regionalization group.

4 Discussion

Overall, all of the regionalization procedures can successfully produce contiguous regions that meet our desired criteria for mapping PHI. Each method can produce regions that: 1) align with census boundaries ; 2) contain populations of at least 20,000 people; and 3) provide a better resolution than the current standard for sharing PHI (3-digit ZCTAs). The regionalizations provided between 99 and 117 units on average per configuration which is far greater than that provided by counties and ZCTAs. Figure 4.7 illustrates how regionalization provides a much higher resolution depiction of the Twin Cities—with around two dozen units which can be used to describe various parts of Minneapolis and St Paul. On the other hand, counties and 3-digit ZCTAs only split the

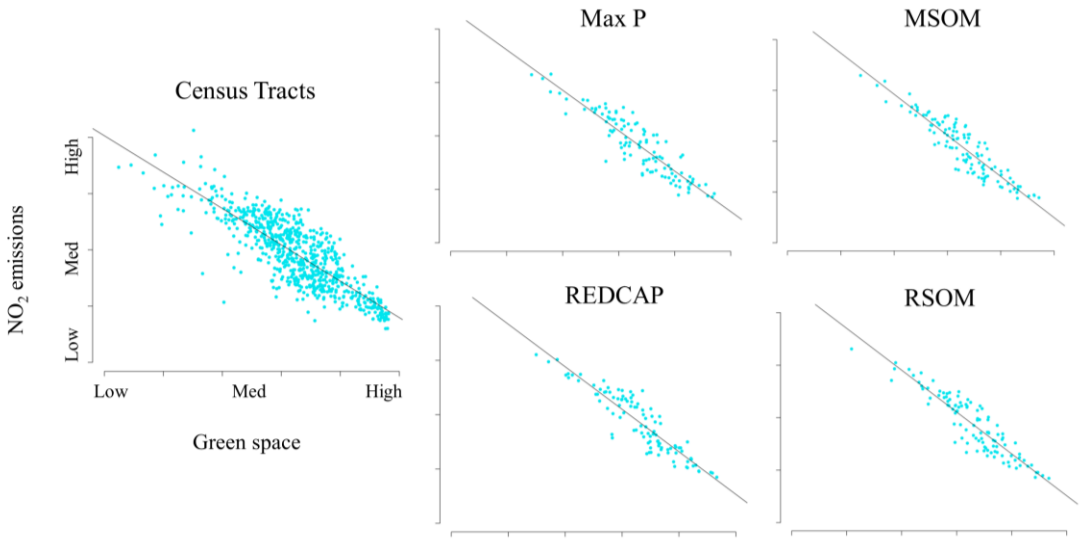
cities into two regions (each containing populations well over 20,000). The oft-used 5-digit zip codes provide a resolution almost as good as what is given by regionalization, but this schema suffers from suppression or holes in the data where some regions are removed because they contain populations less than 20,000. In terms of meeting the 20,000 population threshold required by HIPAA safe harbor, regionalization achieves a full configuration of finer-scaled units for displaying PHI, striking a balance between high resolution and protected data.



4. 7 The Twin Cities (bright blue) shown at the level of 3-digit Zip codes, Counties, and 5-digit Zip codes (aggregations commonly used to share PHI) and compared with regionalization (MSOM in this example). Areas that contain populations under 20,000 are suppressed and are shown in black.

Furthermore, if we were to ignore the HIPAA safe harbor requirements and compare how regionalization fares against finer-scaled (non-HIPAA compliant) alternatives such as census tracts, we find that regionalization offers better fit despite census tracts having a higher resolution (census tracts outnumber regionalization in units by more than 6 to 1). The better relative fit is exemplified in Figure 4.8 which shows how using regionalization

(Max P, MSOM, REDCAP, and RSOM) leads to tighter fitting predictions (smaller average residuals) compared to census tracts in the example of the highly-correlated relationship between greenspace and nitric dioxide exposure. This greater fit is due to how regionalization approaches will optimize partitioning according to median household income and education (which have been shown to covary with greenspace and nitric dioxide exposure). In other words, the optimization process leads to improved groupings of distinct populations, and these distinct populations systematically differ in exposure to green spaces and nitric dioxide. In contrast, census tracts (which are built with socioeconomic homogeneity in mind) are not always homogeneous since populations shift overtime while tract boundaries remain relatively stable. This means that some census tracts contain disparate populations and these tracts would have higher residuals. Therefore, the optimization process in regionalization has implications for how we can better identify communities or neighborhoods. Still, we might argue that although neighborhoods are usually homogenous, some are not (i.e., areas undergoing gentrification), and there is value in maintaining the diversity of these areas within our analysis. This is where other parameters (such as compactness) come into play. Maintaining a certain level of compactness could help to keep some of these diverse neighborhoods intact.



4. 8 A comparison of census tracts versus regionalization (Max P, MSOM, REDCAP, and RSOM) in the example of the highly-correlated relationship between greenspace and nitric dioxide exposure. All five plots are scaled the same on the x-axis and y-axis.

There are many different regionalization strategies and they all differ in terms of how they prioritize the optimization of homogeneity and maintaining compact units. The present study focused on two of the most popular strategies, Max P and REDCAP, and two SOM variants of these procedures. The following paragraphs provide a discussion of the differences observed between these different regionalization approaches in terms of our five assessment measures.

4.2 Global variation among regionalization approaches

There some consistent global differences among the four regionalization approaches, where global is meant in the spatial sense of overall or averaged characteristics that ignore local variation within each regionalization. We continue to focus on the comparisons between the two parents regionalization approaches (Max P and REDCAP) and between parent (non-SOM) and offspring (SOM) approaches. By examining global differences in homogeneity, compactness, resolution, and model-fit among regionalization approaches, we are able to paint a general picture of the advantages and disadvantages of each strategy. We consider how local variations in region traits present across all four strategies in the following section.

4.2.1 Parent vs offspring: How does SOM impact regionalization?

SOM has different effects on Max P and REDCAP. Overall, MSOM was an improved version of Max P for almost every measure. Compared to its parent regionalization (Max P), MSOM provided superior compactness, model-fit (according to the AIC metric), and resolution (in terms of having both smaller average size and less variability between sizes). The only instance in which adding SOM to Max P resulted in significant degradation of region characteristics was for average homogeneity. This means Max P's only advantage was that it maintained the highest average level of homogeneity compared to the other three methods. In contrast, adding SOM to REDCAP did not result in much improvement, whereby RSOM had degraded measures of compactness

and homogeneity, as well as having no significant effects on resolution and AIC. The only improvement we observed was that RSOM had significantly reduced variability of region size compared to REDCAP. For most modeling situations, the single improvement to the variability of region size likely does not outweigh the degradation of compactness and homogeneity. In light of this overall degradation in performance (to mention the extra time and effort it takes to implement the RSOM procedure compared to REDCAP), RSOM may not be a worthwhile approach for many situations.

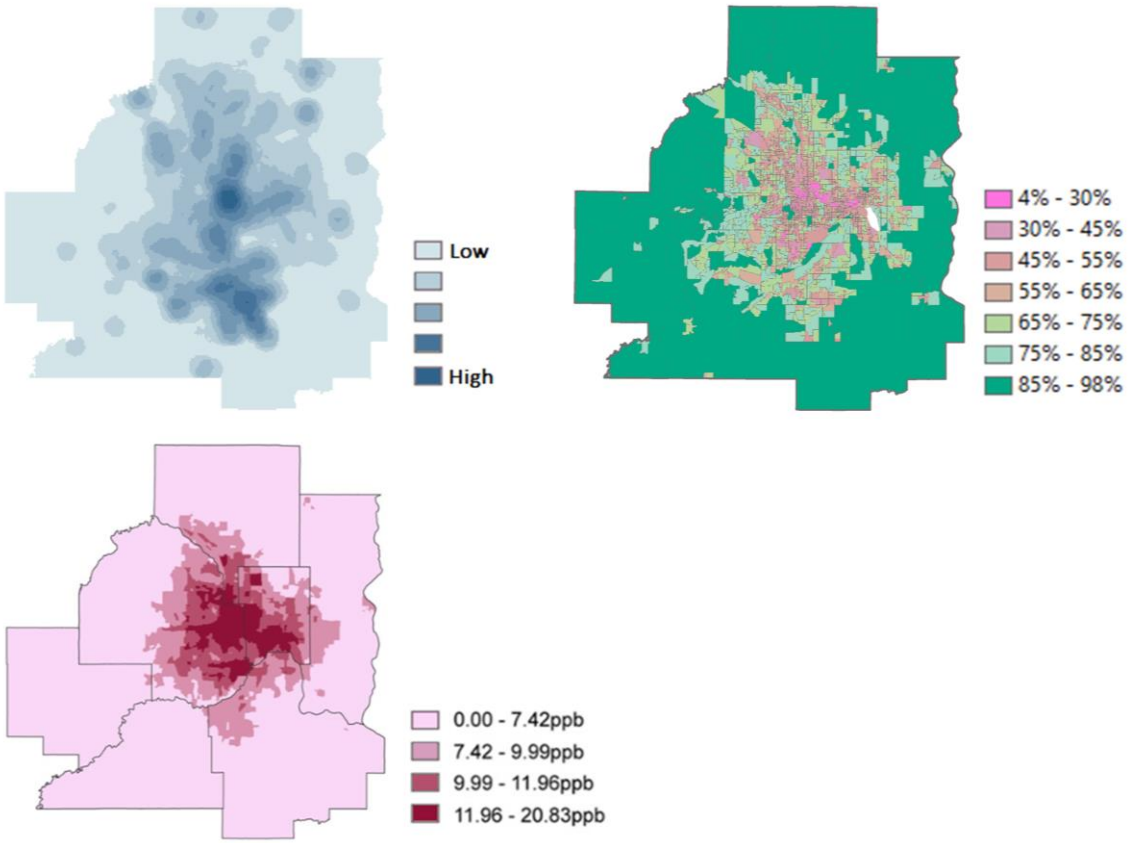
Why does SOM help Max-P but not REDCAP? In essence, SOM impacts region characteristics differently depending on whether the regionalization procedure proceeds in a top-down fashion (hierarchical) or bottom-up (linear integer programming) fashion. Max P and REDCAP delineate regions in different ways and therefore prioritize different characteristics. Max P begins with creating feasible solutions guided by the population threshold and holds off on the optimization of attribute similarity until the final step, while REDCAP does the reverse. REDCAP uses attribute similarity as an initial means to build the spanning tree and then makes cuts guided by attribute similarity and the population threshold in the final step. Another major difference is that Max P integrates a search heuristic (we used tabu search) to test out different arrangements in that final step in order to find the optimal solution (tabu search allows for non-improving moves which dramatically expands the range of solutions tested). REDCAP is simpler in the sense that the range of possible solutions is restricted to one initial spanning tree built in step 1. In other words, Max P creates the (near) optimal solution that prioritizes homogeneity while REDCAP creates the best solution within the reach of its original spanning tree. For these reasons, Max P prioritizes homogeneity while REDCAP indirectly favors compactness by using the hierarchical and nested structure of minimum spanning trees. These procedural differences in regionalization are why SOM ended up improving most of the region characteristics of Max P and degrading most of those of REDCAP.

In terms of specifics of how SOM affects REDCAP and Max P, we can dive into specific measures.

- **Homogeneity**. The SOM procedure essentially smooths the data before inputting it (the u-matrix) into regionalization. Blending the underlying data before regionalization by smoothing population demographics (in this case median household income and education) impacts the output in terms of our region characteristics. The direct degradation of average region homogeneity after adding SOM to both Max P and REDCAP regionalization is straightforward to explain: the smoothing action of SOM degrades the precision of the income variable thereby reducing income-based homogeneity. What is less clear, is by what means SOM impacts the other region characteristics.
- **Resolution**. SOM impacts the variability of the region sizes the same for both Max P and REDCAP, reducing the variability of region sizes. It is difficult to pinpoint with certainty a primary driver of this effect, but it would be reasonable to believe that, by degrading homogeneity in certain spaces, SOM helped to reduce the average region size. By changing the layout of the data (via SOM) we change the tendencies of the aggregation. In this case, changing these tendencies happened to lead to configurations with more, finer units. Additionally, SOM also impacts the average region size of Max P and REDCAP, reducing the average region size (but this reduction is only statistically significant between Max P and MSOM). SOM's impact on resolution is more thoroughly discussed in a later section on the differences in local variation observed *within* each regionalization. This is to say that different areas on the map experienced greater changes to resolution after SOM than others.
- **Compactness.** SOM affects compactness differently for REDCAP than for Max P, whereby adding SOM increased the average compactness of Max P and decreased the average compactness of REDCAP. This is because Max P thrives on the precision of the underlying data, so disrupting the data with SOM reduces homogeneity and results in an indirect improvement in compactness. That is to say, with Max P, when the underlying data is blended, the lowest lows and the highest highs move towards the center of the distribution and the data is made to

be less disparate. This means that, in the final step of Max P regionalization (tabu search), regions are less apt to sprawl because nearest neighbors are made to be more similar and units that are farther away are made to be more disparate. RECAP, on the other hand, makes compactness a priority by the means of its hierarchical structure, so when SOM is applied to RECAP, compactness is degraded due to changes in the initial spanning tree (which would now rely on underlying data that does not reflect homogeneous areas as well). This new spanning tree limits RSOM's ability to make the most of preexisting highly homogenous compact base units. Because this is a factor of the layout of the underlying data, a more thorough discussion is provided in section 4.3 on the local variation.

- **Model-Fit (AIC).** SOM also impacts the model-fit, whereby adding SOM improves the average AIC for Max P but does not improve model-fit for REDCAP to an extent that would be considered statistically significant. It is difficult to pinpoint with certainty a primary driver of this effect, but it would be reasonable to believe that, for Max P, changes in region compactness after SOM may have driven the increased model fit seen in MSOM. The AICs were derived from a model of greenspace and air quality on depression risk, and when we look at a map of depression risk, we notice that it appears to be concentrated in the center of the metro area with a projection of higher risk out east that closely aligns with a projection of low green space and poor air quality (Figure 4.9). It is possible that the heavily homogeneous spaces of Max P capture this model of depression well, and by smoothing our underlying data, SOM helps to capture depression risk even better. This might be due to the increases in compactness seen in the inner city where depression is more heavily concentrated.



4. 9 A map of depression risk in the Twin Cities metro area (top left), green space (top right), and nitric dioxide exposure (bottom left). Depression data was masked using an unspecified smoothing function in addition to having its legend converted to a low-to-high scale.

4.2.2 Parent vs parent: Heuristic vs hierarchical?

In terms of comparing the two (non-SOM) parent strategies, REDCAP on its own provided the highest relative compactness and geosilhouette scores compared to all other regionalization strategies, while offering the second highest average homogeneity. REDCAP seems to strike a good balance between compactness and homogeneity and, for this reason it may potentially provide better representations of communities (by better taking into account populations and space). Results from the geosilhouette models further the notion that REDCAP creates better representations of communities as REDCAP (and its offspring) provided on average much higher geosilhouette scores than Max P. The differences are stark and can be seen in the map and scatter plot in figure 4.6. Max P's overall low average geosilhouette scores can perhaps be somewhat

attributed to Max P's preference for homogeneity over compactness. Max P maintains relatively loose compactness constraints, which allows it to "reach out" and grab similar (yet distant) populations that can be merged together, resulting in a final configuration with elongated or winding regions. This inability to maintain compactness contributes to the observed lower than average geosilhouette scores directly by the means of increasing distance to the next best fit cluster. Distance is a key component in computing path silhouette scores. Per Tobler's law "everything is related to everything else, but near things are more related than distant things" (1970). By using regionalization that values compactness, we acknowledge Tobler's law and step closer to developing more realistic neighborhood units. Regionalization that disregards compactness in favor of homogeneity can result in wonky regions that might appear to be products of gerrymandering or spatial p-hacking rather than suitable representations of neighborhoods.

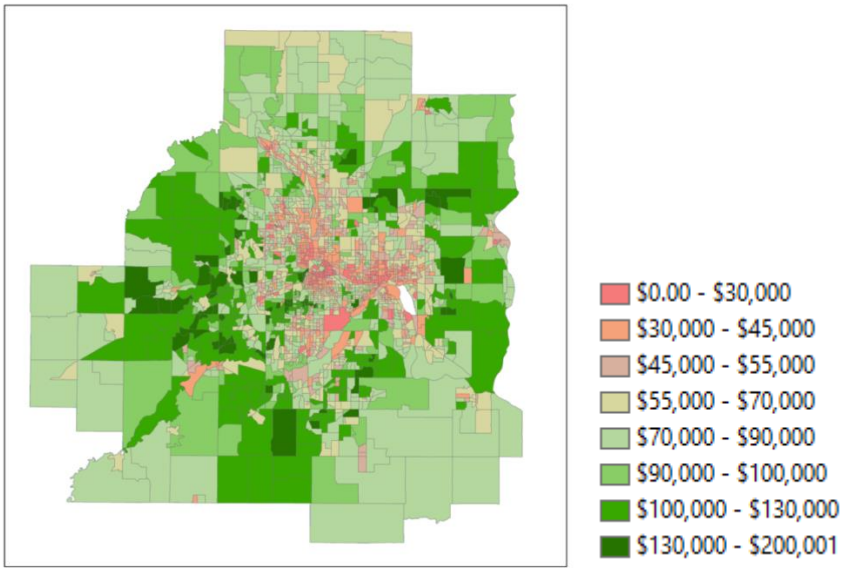
4.3 Local variability

Even though there were clear distinctions between the four regionalization strategies in terms of their average region characteristics, we noticed that behind these means were some very interesting patterns of local variation. The following paragraphs provide deeper insight into the differences in the local distributions of the various region traits.

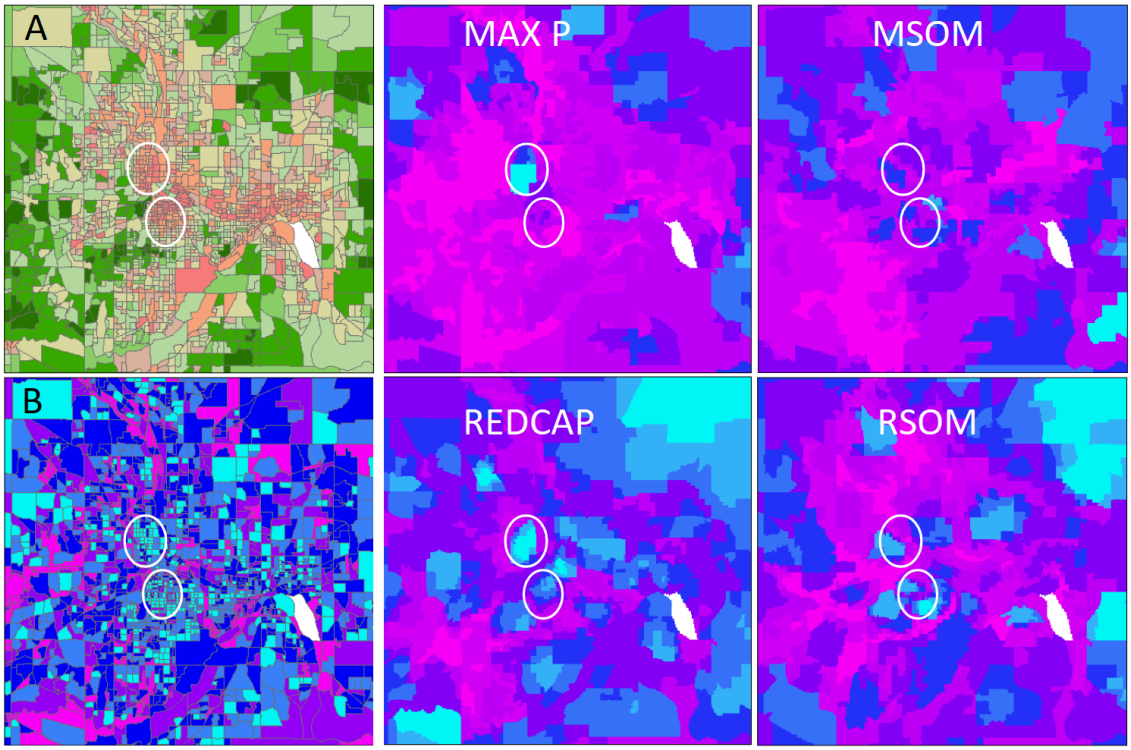
4.3.1 Compactness

When we look at the local variation via the raster average map in figure 4.3, we notice the same thing across the board: less compact regions in the inner metro area and more compact regions in the outer suburbs, with some highly compact regions appearing as spots in the center of the metro (more often with REDCAP and RSOM). This pattern closely follows the layout of the underlying data (the distribution of income is shown in figure 4.10) which exhibits high homogeneity of income and education in the outer suburbs and in smaller spots in the center of the city with low homogeneous areas in between. The pattern of compactness may also be tied to the average size of the building

blocks (which are much smaller in the center of the metropolitan area where population density is relatively larger). Here we would guess that smaller building blocks may tend to provide more freedom of motion by having more boundaries, and therefore more paths, that the regionalization can take which might exacerbate the sprawl of the regions. Max P was more apt to create winding regions in the outer suburbs and missed several opportunities to create highly compact regions from the pre-existing compact homogeneous centers in the suburbs (Waconia, Hugo, and Farmingham) as well as in the spots closer to the city's center (South Minneapolis, the U of M campus, and South Lake Harriet). This is likely a factor of Max P's optimization function working to balance the overall homogeneity for the configuration by sprawling out to achieve higher homogeneity in places that are more diverse in income and education. Because SOM reduces the sprawl of Max P, MSOM was able to capture a greater number of compact homogeneous areas. REDCAP was able to capture more of these pre-existing compact homogeneous spaces (Figure 4.11) than any other method. However, after SOM, the homogeneity of the underlying data was made less precise, and therefore RSOM missed several opportunities to find these pre-existing highly compact homogeneous regions in the center of the metro area which resulted in the global reduced average compactness observed.



4. 10 Median household income of the base units (block groups) for the seven-county metropolitan area.

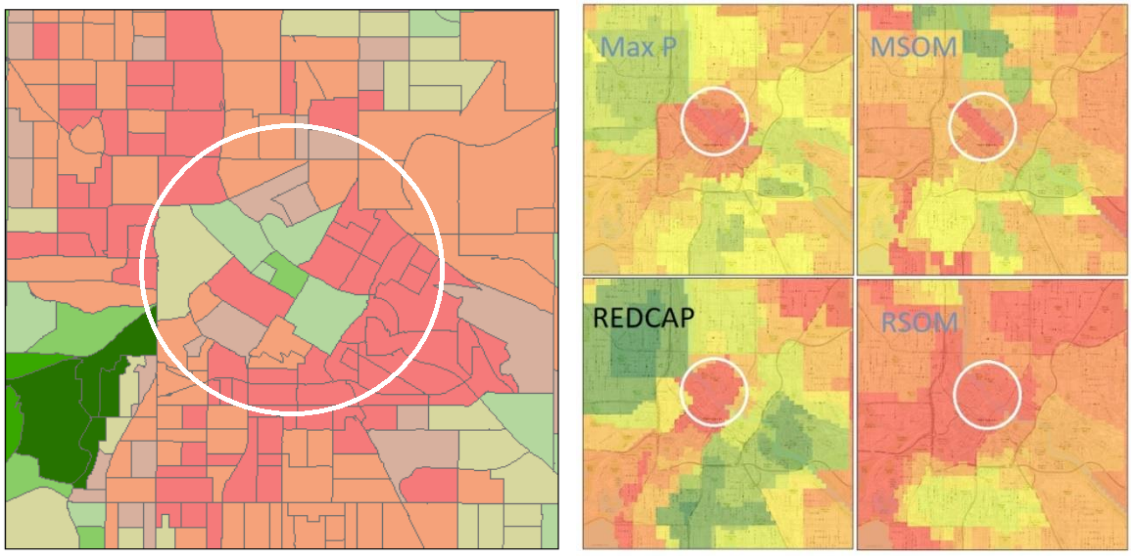


4. 11 A) Median household income of the base units (block groups). B) Compactness of the base units. Circled are two examples of highly compact and homogeneous spaces (North Minneapolis and South Minneapolis) that exist within the data at the base unit level. REDCAP and RSOM do a better job of finding these two spaces in addition to providing a number of moderately compact (dark blue) spots in the inner city area, while Max P only finds North Minneapolis.

4.3.2 Homogeneity.

When we look at local variation as seen from the raster maps in figure 4.4 we notice less homogeneous regions in the inner metro area and more homogenous in the outer suburbs. The northwest quadrant of the maps (around Rogers and Ramsey) has the highest homogeneity for all four approaches. This may be a factor of this area having some of the least diversity in terms of income and education. It is also the case that all four approaches provided low homogeneity in the Warehouse District where there is a clustering of block groups characterized by high income and high education that, when combined, have a population that does not meet the minimum threshold which means these units must be lumped with neighboring units. Because the surrounding units are those of low income and low education it is seemingly impossible for any of the four strategies to create a homogenous region in this center (Figure 4.12). In terms of the

differences observed in the local variations among strategies we notice that, even though Max P had the highest average homogeneity, REDCAP seems to do a better job at maintaining high homogeneity within the city's center in North Minneapolis, the U of M, and South Minneapolis, which is easily seen when comparing the green areas between Max P and REDCAP in the figure below. This is a factor of Max P's tendency to sprawl as the areas just mentioned are made up of relatively compact base units.

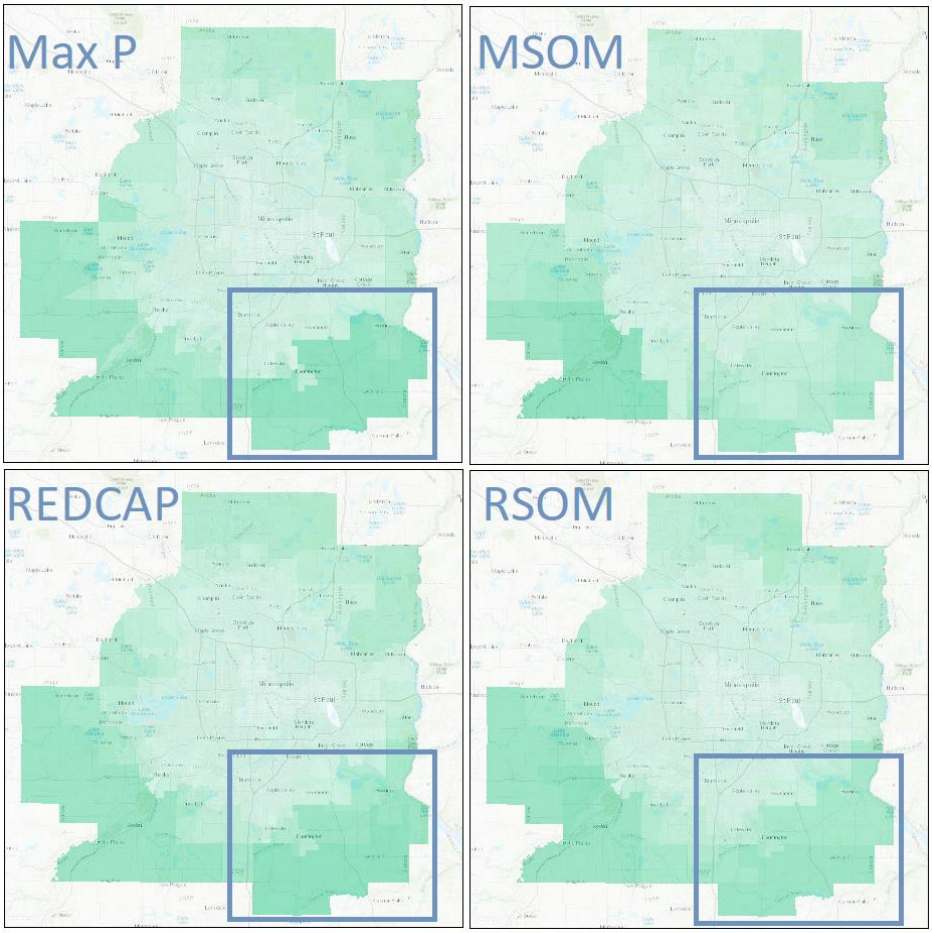


4. 12 A cluster of five high income block groups in the Warehouse District surrounded by low income units (left). These units have a combined total population of less than 20,000 people and therefore all four regionalization strategies fail to create a homogeneous region in this area (right).

4.3.3 Resolution

When considering local variation we notice the smaller regions in the inner city and larger in the outer suburbs. This is what we'd expect given that the size of our input units (block groups) varies by population density. Smoothing the data with SOM enabled regionalization to aggregate units into smaller regions for certain areas of the map. For example, the SE quadrant of the metro area experienced a noticeable reduction in region size after applying SOM. By reducing the size of the regions in this corner (which had the largest region sizes in Max P and REDCAP) we cut out the high end of the range and reduced the overall variability in region sizes. Furthermore, we notice that this reduction

was more noticeable in MSOM than RSOM (Figure 4.13). For this reason, we believe that SOM's action in the SE quadrant is what drove the significant improvement in resolution that we observed when comparing MSOM to Max P. RSOM's reduction of region size in the SE corner was subtle and, although it reduced the variability in size—it was not enough to bring down the average region size to a statistically significant degree.



4. 13 A cluster of regions in the south east metropolitan region with the highest average region size for parent regionalizations is reduced after the addition of SOM. This reduction is more apparent when comparing Max P to MSOM.

4.3.4 Geosilhouettes.

In terms of the local variation of geosilhouette scores, we notice very different patterns of geosilhouette scores between the two parent/offspring regionalization pairs (Figure 4.6). Here we see that even though adding SOM to Max P resulted in a slight (but not statistically significant) decrease in geosilhouette scores, it did not decrease the scores

evenly across all regions. For MSOM, it seems that scores worsened in the lower southeast quadrant, the north, and the northwest, with not too much change to the center of the map. These areas of change are characterized by smaller farm towns beyond the outlying suburbs. Adding SOM to REDCAP did not result in noticeable local variation—but for perhaps a slight worsening in the southeastern quadrant. In terms of the statistically significant difference in geosilhouette scores between Max P and REDCAP, we could easily imagine that REDCAP's dramatically higher average geosilhouette scores comes from its greater relative compactness. Geosilhouette scores are computed using a path dissimilarity metric which could theoretically penalize elongated regions by increasing the distance of, and difference observed along, the path that separates a block group and its next best connected region. This would also explain why we see a slight decrease (which was not statistically significant) in geosilhouette scores when going from REDCAP to RSOM (which has less compact regions than REDCAP). The quandary is that we observed a relative decrease (which was not statistically significant) in geosilhouette scores when going from Max P to MSOM even though MSOM is significantly more compact than Max P. This would suggest that there is something more than compactness at play. Geosilhouette scores are composed of both spatial and attribute similarity metrics which means that the homogeneity of the area plays an important role as well. MSOM's average homogeneity is dramatically less than that of Max P in terms of income, and income has been tied to depression risk (Patel et al., 2018). For this reason, it is not outside of the realm of possibility that Max P would have higher homogeneity in terms of depression risk as well. Compactness is likely the primary driver of the results (as exemplified by the differences between Max P and REDCAP) however homogeneity might be the reason for the reduction of scores observed in MSOM.

5 Conclusion

There is a real need for finding ways to work with and share neighborhood-level health data. This project addresses this need by presenting regionalization as a strategy for creating custom fine-scaled units for sharing PHI without breaching patient privacy regulations. Without regionalization, investigators would continue to rely on the present

standard for sharing PHI, the county or ZCTA, which has repeatedly been deemed insufficient for neighborhood-level studies of health. By sharing data at finer resolutions and in more meaningful forms than ZCTAs, we provide more accurate depictions of neighborhood health. This project explores four different regionalization strategies, each having its own strengths and weaknesses in terms of the neighborhood configuration it creates. Investigators are encouraged to use the strategy that best suits the needs of the project to be visualized and shared, however, the current project showed that REDCAP proves to be a superior approach to regionalization for the analysis and display of PHI, providing relatively high scores on characteristics most important for neighborhood health (compactness, homogeneity, and model-fit), as well as providing much finer regions than the standard approaches we rely on today. Additionally, MSOM—which provided the finest grained units—stands to offer an improved version of Max P for those who require a bottom-up procedure or can't access REDCAP.

In terms of limitations and future research considerations, our results were conditioned by a number of choices relating to data. First, we relied on regionalization methods that optimized on median household income and education (proportion of population with bachelor's degrees or higher). The arrangement of income and education weighs heavily on the performance of the regionalization (especially for Max P regions), and therefore it is unclear whether our results would hold true for other study sites with different spatial arrangements of income and education. Second, the results obtained from this study may not generalize across health outcomes. For instance, it is possible that different results could have been obtained for our model-fit metrics if we used a different disease (infectious disease instead of depression). Third, like many cases studies, ours is affected by the *boundary problem*; that is, the units at the edge of the study site have fewer neighbors compared to inside units, which means some aggregations almost necessarily end up having the same set of blocks. It is possible that the boundary problem may have impacted some or all of the region characteristics including compactness, homogeneity, and resolution. In order to avoid this problem, future studies could carry out regionalization while including the units from surrounding counties and then clipping the regions to the seven-county metro area at a final step. Given these potential limitations, future research could focus on the impacts of a different data set or study area. For

example, it would be interesting to repeat the study using other health and disease outcomes, within other metropolitan areas in the US, or within various sets of simulated data—perhaps with varying degrees of spatial autocorrelation.

The study also reflects choices around methods. First, we explored the use of SOM to smooth multidimensional data to be inputted into two different parent regionalization procedures. It is possible that other (simpler) smoothing methods than SOM may serve the same purpose. Future research exploring the impacts of preprocessing input data with various smoothing methods and comparing and contrasting these outputs with those from MSOM might provide more insight into how to improve Max P regionalization. Second, since our focus was on assessing the use of regionalization for PHI, we chose the most commonly used or standard setting for our methods, but we recognize that there is room to experiment with modifying any of our chosen approaches. With our exploration of REDCAP in particular, we examined just one of the six approaches under the REDCAP banner. We chose average-linkage clustering with full-order constraining which is the approach that offers to provide the greatest number of units as determined by Kugler et al (2017). Other families may provide different results—especially given that the SOM effect was heavily dependent on the regionalization procedure. Full-order single-linkage, for example, has been shown to outperform average-linkage in certain scenarios and therefore might improve REDCAP resolution without severely degrading compactness (Kugler et al, 2017). In sum, future research into the other families of REDCAP would be advantageous.

Chapter 5. Conclusions and Future Directions

Over the course of three papers, this dissertation has shed light on the scarcity of maps and spatial analyses published within the literature on neighborhood health and pointed to a potential cause being the ambiguous rules that guide how researchers can share geographic data. By providing a thorough examination of the safe harbor provision specific to geographic data, this dissertation helped elucidate the ambiguity within the law to encourage safe and effective sharing of mapped patient data. Finally, this dissertation also presented a number of regionalization strategies that offer to help investigators work within the bounds of privacy provisions to share maps and spatial data.

1 Understanding the value of maps in public health.

Although this project hoped to undercover easily resolvable barriers to sharing maps and spatial analyses, the survey results indicated a more complex barrier stands in the way. The primary barrier identified by survey respondents was the belief that a map would not add further insight beyond that of which was provided by statistical models. This is to say that many neighborhood health investigators did not see the value in spatial data visualization. Unlike other barriers, such as time constraints or lack of resources which can be resolved by updating technical and teaching tools, the belief that maps do not add value to neighborhood health research is more difficult to address. Future research should attempt to gain a better understanding of the specific ways in which geovisualization is valued, or not valued. This kind of research can help further awareness while at the same time demonstrate the value of spatial data visualization by providing illustrative examples of the advantages of maps in neighborhood health. Over time, as maps and spatial analyses appear more frequently within the literature, the success of these studies will help encourage others to follow suit. With luck, sometime in the near future spatial data exploration will become common practice in neighborhood health research.

2 Having separate privacy regulations for maps and tables.

Despite ongoing examples of misinterpretation, the safe harbor rule stands as the primary guidance for those interested in sharing maps. This privacy provision is overly conservative and alternative methods have potential to do a better job at sharing protected health information in ways that keep the data useful and safe. However, with rapidly evolving technology and the amount of individual-level data collected by companies ever increasing, it becomes more and more difficult to foresee policy makers comfortably loosening data protection guidelines. One way forward might be to focus on the differences between the level of information shared within tables linked to aggregated map units and tables of individual-level microdata. It is not possible for aggregated map units to provide individual level information such as gender or birth date and therefore map tables are much less vulnerable to the dangers of identity attacks from linked tables. For this reason, it is foreseeably possible to loosen the geographic constraints of the safe harbor rule in instances where aggregated geocodes and aggregated risks, and nothing else, are shared. Future research should focus on determining the identification risk involved when sharing aggregated geocodes of 20,000 people or greater. Although it is not likely to find a one-size-fits-all strategy, it is possible that aggregations of 20,000 people may provide sufficient data protection in most instances.

3 Augmenting regionalization and finding better evaluation methods.

If it were determined to be acceptable under the safe harbor rule to share aggregated geocodes with populations of 20,000 people, our results offer four different easy-to-use methods that can help researchers design finer-grained units for displaying and sharing mapped health data. The present study recommends two of these approaches in particular (REDCAP and MSOM). Going forward, due to the relative better performance of REDCAP on a number of measures, it would be advantageous to consider testing the other REDCAP families and assessing the SOM-variants of each. SOM, and perhaps even other smoothing methods, should be further explored as they offer to potentially

improve compactness, resolution, and model-fit for bottom-up regionalization processes. Improvements to these region characteristics could be useful for more than just the display and analysis of patient data. Regionalization in particular has the potential to be useful for helping maintain compactness in the context of political redistricting—a problem that researchers have been grappling with for many years. Further research into finding ways to augment regionalization processes could help a broad array of domains trying to tackle zoning problems which can be computationally intensive when managing a vast number of units over a large spatial extent. Additionally, a greater amount of time and effort should go into finding stronger ways to evaluate neighborhood representation. The current project offers a broader range of measures and serves as a proof of concept for two recently offered methods that have specific advantages for use in regionalization studies (Pinzari's homogeneity index and Wolf et al's geosilhouettes). Research should focus on continuing the development of new and innovative assessment measures that integrate space because these kinds of measures are highly valuable to the field of neighborhood health.

There are several ways in which future research can build upon the various threads of the research discussed in this three-paper dissertation. First, investigators should continue to examine the state of the literature on neighborhood health and push for a spatially aware research agenda until the use of mapping and spatial analysis becomes common practice. Second, there is a need to assess the extent to which maps can be safely shared when geocodes are aggregated to contain populations of at least 20,000 people; this could help lead to the development of separate regulations for maps and microdata. Finally, investigators should explore ways in which regionalization can be augmented with SOM and other data smoothing methods so to make it more useful for research in public health and a variety of other domains.

Complete Dissertation References

Acevedo-Garcia, D. (2001). Zip code-level risk factors for tuberculosis: neighborhood environment and residential segregation in New Jersey, 1985-1992. *American journal of public health*, *91*(5), 734. https://doi.org/10.2105/AJPH.91.5.734

Anders, M. E., & Evans, D. P. (2010). Comparison of PubMed and Google Scholar literature searches. *Respiratory care*, *55*(5), 578-583.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American statistician*, 27(1), 17–21.

Auchincloss, A. H., Gebreab, S. Y., Mair, C., & Diez Roux, A. V. (2012). A review of spatial methods in epidemiology, 2000–2010. *Annual review of public health*, *33*, 107-122.

Bação, F., Lobo, V., & Painho, M. (2004, October). Geo-self-organizing map (Geo-SOM) for building and exploring homogeneous regions. In *International Conference on Geographic Information Science* (pp. 22-37). Springer, Berlin, Heidelberg.

Bação, F., Lobo, V., & Painho, M. (2005, May). Self-organizing maps as substitutes for K-means clustering. In *International Conference on Computational Science* (pp. 476-483). Springer, Berlin, Heidelberg.

Barth-Jones, D. (2012). The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *SSRN Electronic Journal*, 1-19. https://doi.org/10.2139/ssrn.2076397

Beall, C. M. (1983) Ages at menopause and menarche in a high-altitude Himalayan population, *Annals of Human Biology*, 10:4, 365-370, DOI: 10.1080/03014468300006531

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, *17*(2), 169–177. https://doi.org/10.1136/jamia.2009.000026

Best, S. J., Krueger, B. S., & Ladewig, J. (2006). Privacy in the information age. *International Journal of Public Opinion Quarterly*, *70*(3), 375-401. https://doi.org/10.1093/poq/nfl018

Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. Annual Review of Public Health, 25, 53–77.

Browne, A. C., Kayaalp, M., Dodd, Z. A., Sagan, P., & McDonald, C. J. (2014). The challenges of creating a gold standard for de-identification research. In *AMIA Annual Symposium Proceedings* (Vol. 2014, p. 353). American Medical Informatics Association.

Burnham, K. P., & Anderson, D. R. (2004). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach* (second). Springer, New York, NY.

Cabrera-Barona, P., Wei, C., & Hagenlocher, M. (2016). Multiscale evaluation of an urban deprivation index: Implications for quality of life and healthcare accessibility planning. *Applied Geography*, *70*, 1-10. https://doi.org/10.1016/j.apgeog.2016.02.009

Cakmak, S., Mahmud, M., Grgicak-Mannion, A., & Dales, R. E. (2012). The influence of neighborhood traffic density on the respiratory health of elementary schoolchildren. *Environment International*, *39*(1), 128-132.

Chaix, B. (2009). Geographic life environments and coronary heart disease: a literature review, theoretical contributions, methodological updates, and a research agenda. *Annual review of public health*, *30*, 81-105.

Ciriani, V., Di Vimercati, S. D. C., Foresti, S., & Samarati, P. (2007). Microdata protection. In *Secure data management in decentralized systems* (pp. 291-321). Springer, Boston, MA.

Croft, W. L., Shi, W., Sack, J. R., & Corriveau, J. P. (2016). Location-based anonymization: comparison and evaluation of the Voronoi-based aggregation system. *International Journal of Geographical Information Science*, *30*(11), 2253-2275.

Curtis, A. (2008). From Healthy Start to Hurricane Katrina: Using GIS to eliminate disparities in perinatal health. *Statistics in Medicine*, *27*, 3984–3997. https://doi.org/10.1002/sim

Curtis, A., Mills, J. W., Agustin, L., & Cockburn, M. (2011). Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems*, *35*(1), 57–64. https://doi.org/10.1016/j.compenvurbsys.2010.08.002

Dao, T. H. D., & Thill, J. C. (2018). Detecting attribute-based homogeneous patches using spatial clustering: a comparison test. In *Information Fusion and Intelligent Geographic Information Systems (IF&IGIS'17)* (pp. 37-54). Springer, Cham.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, *16*(4), 309-330.

Diez Roux, A. V. (2004). Estimating neighborhood health effects: the challenges of causal inference in a complex world. *Social Science and Medicine*, 58(10), 1953-60.

Diez Roux, A. V., & and Mair, C. (2010). Neighborhoods and health. Annals of the New York Academy of Sciences, 1186(1),125-145.

Dodge, S. (2021). A data science framework for movement. *Geographical Analysis*, *53*(1), 92-112.

Duque, J. C., Dev, B., Betancourt, A., & Franco, J. L. (2011). ClusterPy: Library of spatially constrained clustering algorithms, version 2.7.10.

Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming* (pp. 1-12). Springer, Berlin, Heidelberg.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–487. https://doi.org/10.1561/0400000042

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Preserving statistical validity in adaptive data analysis. *Proceedings of the Annual ACM Symposium on Theory of Computing*, 117–126. https://doi.org/10.1145/2746539.2746580

Emslie, C., & Mitchell, R. (2009). Are there gender differences in the geography of alcohol-related mortality in Scotland? An ecological study. *BMC Public Health*, *9(1)*, 1–8.

Entwisle, B. (2007). Putting people into place. *Demography*, 44(4), 687–703. https://doi.org/10.1353/dem.2007.0045

Federal Committee on Statistical Methodology. (1994). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology.

Fei, X., Lou, Z., Christakos, G., Liu, Q., Ren, Y., & Wu, J. (2016). A geographic analysis about the spatiotemporal pattern of breast cancer in Hangzhou from 2008 to 2012. *PLoS One*, 11(1), 1-13.

Freedman, V. A., Grafova, I. B., & Rogowski, J. (2011). Neighborhoods and chronic disease onset in later life. *American Journal of Public Health*, 101(1), 79-86.

Gerber, Y., Weston, S., Killian, J., Therneau, T., Jacobsen, S., & Roger, V. (2008). Neighborhood Income and Individual Education: Effect on Survival After Myocardial Infarction. *Mayo Clin Proc*, *83*(6), 663–669.

Goodchild, M. F. (2011). Scale in GIS: An overview. *Geomorphology*, *130*(1–2), 5–9. https://doi.org/10.1016/j.geomorph.2010.10.004

Greenberg, B., & Voshell, L. (1990). The geographic component of disclosure risk for microdata. In *Statistical Research Division Report Series Census/SRD/RR-90/13, US Bureau of the Census.*

Gu, D., Zhu, H., & Wen, M. (2015). Neighborhood-health links: Differences between rural-to-urban migrants and natives in Shanghai. *Demographic Research, 33(1),* 499-524.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, *22*(7), 801-823. https://doi.org/10.1080/13658810701674970

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS ONE*, *10*(9), 1-17.

Hallowell, B. D., Robb, S. W., & Kintziger, K. W. (2018). Comparing the geographic distribution and location characteristics of HIV-seropositive and HIV-seronegative individuals with a diagnosis of cancer living in the southeast US. *Spatial and spatio-temporal epidemiology*, *24*, 11-18.

Bosma, H., Dike Van De Mheen, H., Borsboom, G. J., & Mackenbach, J. P. (2001). Neighborhood socioeconomic status and all-cause mortality. *American Journal of Epidemiology*, *153*(4), 363-371.

Health Coverage Availability and Affordability Act, *H.R.* 3103,104th Congress, 2nd Session. (1996).

Horm, J. (2000). A Simulation Study of the Identifiability of Survey Respondents when their Community of Residence is Known. *National Center for Health Statistics*.

Iroh Tam, P. Y., Krzyzanowski, B., Oakes, J. M., Kne, L., & Manson, S. (2017). Spatial variation of pneumonia hospitalization risk in Twin Cities metro area, Minnesota. *Epidemiology and Infection*. 145(15),3274-3283. https://doi.org/10.1017/S0950268817002291

Jacquez, G. M. (2000). Spatial analysis in epidemiology: Nascent science or a failure of GIS? *Journal of Geographical Systems*, *2*(1), 91–97.

Janmey, V., & Elkin, P. L. (2018). Re-identification risk in HIPAA de-identified datasets: The MVA attack. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 1329-1337.

Jung, H. W., & El Emam, K. (2014). A linear programming model for preserving privacy when disclosing patient spatial information for secondary purposes. *International Journal of Health Geographics*, *13*(1), 1–6. https://doi.org/10.1186/1476-072X-13-16

Krzyzanowski, B., Manson, S. M., Eder, M. M., Kne, L., Oldenburg, N., Peterson, K., ... & Duval, S. (2019). Use of a Geographic Information System to create treatment groups for group-randomized community trials: The Minnesota Heart Health Program. *Trials, 20*(1), 1-7.

Kugler, T. A., Manson, S. M., & Donato, J. R. (2017). Spatiotemporal aggregation for temporally extensive international microdata. *Computers, Environment and Urban Systems, 63*, 26-37. https://doi.org/10.1016/j.physbeh.2017.03.040

Kwok, P., Davern, M., Hair, E., & Lafky, D. (2011). Harder than you think: a case study of re-identification risk of HIPAA-compliant records. *Proceedings of the 2011 Joint Statistical Meetings.* Chicago: NORC at The University of Chicago. Abstract, 302255.

Li, W., Church, R. L., & Goodchild, M. F. (2014). The p-compact-regions problem. *Geographical Analysis*, *46*(3), 250–273. https://doi.org/10.1111/gean.12038

Liu, W., Yang, K., Qi, X., Xu, K., Ji, H., Ai, J., … Zhu, Y. (2013). Spatial and temporal analysis of human infection with avian influenza A(H7N9) virus in China, 2013. *Eurosurveillance*, *18*(47), 1–8. https://doi.org/10.2807/1560-7917.ES2013.18.47.20640

Malin, B., Benitez, K., & Masys, D. (2011). Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*, *18*(1), 3-10. https://doi.org/10.1016/j.physbeh.2017.03.040

Mitchell, R. J., Richardson, E. A., Shortt, N. K., & Pearce, J. R. (2015). Neighborhood environments and socioeconomic inequalities in mental well-being. *American journal of preventive medicine*, *49*(1), 80-84.

Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; HHS Rules and Regulations, Vol. 78, No.17 Fed. Reg. 5566 (January 25, 2013) (to be codified at 45 C.F.R. pts. 160 and 164 Rin:0945-AA03.

Mu, L., Wang, F., Chen, V. W., & Wu, X. C. (2015). A place-oriented, mixed-level regionalization method for constructing geographic areas in health data dissemination and analysis. *Annals of the Association of American Geographers*, *105*(1), 48-66.

Muralidhar, K., Domingo-Ferrer, J., & Martínez, S. (2020, September). E-Differential Privacy for Microdata Releases Does Not Guarantee Confidentiality (Let Alone Utility). In *International Conference on Privacy in Statistical Databases* (pp. 21-31). Springer, Cham.

Nandi, A., & Harper, S. (2015). How consequential is social epidemiology? A review of recent evidence. *Current Epidemiology Reports*, *2*(1), 61-70.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings IEEE Symposium on Security and Privacy*, 111–125. https://doi.org/10.1109/SP.2008.33

National Committee on Vital and Health Statistics. (Dec 19, 2007). Enhanced Protections for Uses of Health Data: A Stewardship Framework for 'Secondary Uses' of Electronically Collected and Transmitted Health Data. *Report to the Secretary of the U.S. Department of Health and Human Services.* Http://www.ncvhs.hhs.gov/071221lt.pdf.

Nicholson, S., & Smith, C. A. (2007). Using lessons from health care to protect the privacy of library users: Guidelines for the de-identification of library data based on HIPAA. *Journal of the American Society for Information Science and Technology*, *58*(8), 1198–1206. https://doi.org/10.1002/asi.20600

O'Neill, L., Dexter, F., & Zhang, N. (2016). The risks to patient privacy from publishing data from clinical anesthesia studies. *Anesthesia & Analgesia*, *122*(6), 2017-2027.

Oakes, J. M. (2004). The (mis) estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social Science & Medicine*, *58*(10), 1929-1952.

Oakes, J. M., Andrade, K. E., Biyoow, I. M., & Cowan, L. T. (2015). Twenty years of neighborhood effect research: an assessment. *Current Epidemiology Reports*, *2*(1), 80-87.

Oberski, D. L., & Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1), 1-21.

Office for Civil Rights. (March 28, 2017). Workshop on the HIPAA privacy rule's de-identification standard. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/2010-de-identification-workshop/index.html

Office for Civil Rights. (November 26, 2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, *2*(4), 459–472. https://doi.org/10.2307/622300

Openshaw, S. (1983). The modifiable area unit problem. *Concepts and Techniques in Modern Geography*, *38*, 1–41.

Osserman, R. (1978). The isoperimetric inequality. *Bulletin of the American Mathematical Society*, *84*(6), 1182-1238.

Page, S. E. (2008). Agent-based Models. In S. N. Durlauf and L. E. Blume (Eds). *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Patel, V., Burns, J. K., Dhingra, M., Tarver, L., Kohrt, B. A., & Lund, C. (2018). Income inequality and depression: a systematic review and meta-analysis of the association and a scoping review of mechanisms. *World Psychiatry*, *17*(1), 76-89.

Pinzari, L., Mazumdar, S., & Girosi, F. (2018). A framework for the identification and classification of homogeneous socioeconomic areas in the analysis of health care variation. *International Journal of Health Geographics*, *17*(1), 1-17.

Relia, K., Akbari, M., Duncan, D., & Chunara, R. (2018). Socio-spatial self-organizing maps: using social media to assess relevant geographies for exposure to social processes. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1-23.

Representative Gingrich. "Conference Report on H.R. 3845, District of Columbia Appropriations Act, 1997." Congressional Record (August 1, 1996) p. H9801. Available from: LexisNexis® Congressional; Accessed: 9/19/20.

Roblin, D. W. (2013). Validation of a neighborhood SES index in a managed care organization. *Medical Care*, *51*(1), 1-8.

Rose, A. N., & Nagle, N. N. (2017). Validation of spatiodemographic estimates produced through data fusion of small area census records and household microdata. *Computers, Environment and Urban Systems*, *63*, 38-49.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019, May). Differential privacy and census data: Implications for social and economic research. In *AEA Papers and Proceedings*, 109,403-08.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *the Symposium on Principles of Database Systems (PODS98)*, 98(188),10-1145.

Santos-Lozada, A. R., Howard, J. T., & Verdery, A. M. (2020). How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences*, *117*(24), 13405-13412.

She, B., Duque, J. C., & Ye, X. (2017). The network-max-P-regions model. *International Journal of Geographical Information Science*, *31*(5), 962-981.

Sheu-jen, H., Wen-chi, H., Patricia, A. S., & Jackson, P. W. (2010). Neighborhood environment and physical activity among urban and rural schoolchildren in Taiwan. *Health & Place*, *16*(3), 470-476.

Siahpush, M., Heller, G., & Singh, G. (2005). Lower levels of occupation, income and education are strongly associated with a longer smoking duration: multivariate results from the 2001 Australian National Drug Strategy Survey. *Public Health*, *119*(12), 1105-1110.

Spielman, S. E., & Folch, D. C. (2015). Reducing uncertainty in the American Community Survey through data-driven regionalization. *PloS ONE, 10*(2), 1-21.

Standards for Privacy of Individually Identifiable Health Information; HHS Rules and Regulations, Vol. 65, No.250 Fed. Reg. 82462 (December 28, 2000)(to be codified at 45 C.F.R. pts. 160 and 164 Rin:0991-AB08.

Standards for Privacy of Individually Identifiable Health Information; Proposed Rules, 64 No. 212 Fed. Reg. 59918 (November 3, 1999) (to be codified at 45 C.F.R. pts. 160 through 164 RIN 0991-AB08.

Standards for Privacy of Individually Identifiable Health Information; Final Rule, 65 No. 250 Fed. Reg. 82462 (December 28, 2000) (to be codified at 45 C.F.R. pts. 160 through 164 RIN 0991-AB08.

Standards for Privacy of Individually Identifiable Health Information; Final Rule, 66 No. 38 Fed. Reg. 12434 (February 26, 2001) (to be codified at 45 C.F.R. pts. 160 through 164 RIN 0991-AB08.

Standards for Privacy of Individually Identifiable Health Information; Final Rule; request for comments, 66 No. 40 Fed. Reg. 12738 (February 28, 2001) (to be codified at 45 C.F.R. pts. 160 through 164 RIN 0991-AB08.

Stansfeld, S., Haines, M., & Brown, B. (2000). Noise and health in the urban environment. *Reviews on Environmental Health*, *15*(1-2), 43-82.

Subramanian SV. (2004). The relevance of multilevel statistical methods for identifying causal neighborhood effects: Comment. Social Science and Medicine, 58(10):1961-7.

Sweeney, L. (1997). Guaranteeing anonymity when sharing medical data, the Datafly System. In *Proceedings of the AMIA Annual Fall Symposium* (p. 51). American Medical Informatics Association.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3. Carnegie Mellon University, Pittsburgh.

Sweeney, Latanya, Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., & Brody, J. G. (2017). Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. *Technology Science*, *2017*, 1–75.

Tellman, N., Litt, E. R., Knapp, C., Eagan, A., Cheng, J., & Lewis Jr, J. (2010). The effects of the Health Insurance Portability and Accountability Act privacy rule on influenza research using geographical information systems. *Geospatial Health*, *5*(1), 3-9.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg.

Van Os, J., Hanssen, M., Bijl, R. V., & Vollebergh, W. (2001). Prevalence of psychotic disorder and community level of psychotic symptoms: an urban-rural comparison. *Archives of General Psychiatry*, *58*(7), 663-668.

Van Wijk, J. J. (2005). The value of visualization. In *VIS 05. IEEE Visualization*, 79-86. 10.1109/VISUAL.2005.1532781

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *portal: Libraries and the Academy*, 9(1), 5-24.

Wang, F., Guo, D., & McLafferty, S. (2012). Constructing geographic areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois. *Applied Geography*, *35*(1-2), 1-11.

Wolf, L. J., Knaap, E., & Rey, S. (2021). Geosilhouettes: Geographical measures of cluster fit. *Urban Analytics and City Science*, *48*(3), 521-539.

Yu, L., Recker, M., Chen, S., Zhao, N., & Yang, Q. (2018). The moderating effect of geographic area on the relationship between age, gender, and information and communication technology literacy and problematic internet use. *Cyberpsychology, Behavior, and Social Networking*, *21*(6), 367-373.

Zayatz, L. V. (1992). Estimation of the number of unique population elements using a sample. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 369–73.

Zhang, X., Holt, J. B., Lu, H., Onufrak, S., Yang, J., French, S. P., & Sui, D. Z. (2014). Neighborhood commuting environment and obesity in the United States: An urban-rural stratified multilevel analysis. *Preventive Medicine*, 59(1), 31–36.

Appendix

Paper 1 article list

Author/Year	Journal	*contains map
*Acevedo-Garcia et al., 2001	American Journal of Public Health	
Airaksinen et al., 2016	European Journal of Public Health	
*Alcaraz et al., 2009	Preventive Medicine	
Auchincloss et al., 2007	Epidemiology	
Auchincloss et al., 2013	Obesity	
Auchincloss et al., 2001	Journal of Aging and Health	
*Barr et al., 2001	American Journal of Public Health	
Beard et al., 2009	American Journal of Public Health	
Beck et al., 2015	JAMA Pediatrics	
Beck et al., 2013	Journal of Pediatrics	
Beck et al., 2017	Journal of Urban Health	
Berke et al., 2010	American Journal of Public Health	
*Beyer et al., 2014	Int. Journal of Environmental Research Public Health	
Bluthenthal et al., 2008	Journal of Urban Health	
Boardman, 2004	Social Science and Medicine	
Boardman et al., 2001	Journal of Health and Social Behavior	
Borrell et al., 2011	Ethnicity and Disease	
Borrell et al., 2004	Community Dentistry and Oral Epidemiology	
Bostean et al., 2018	Health and Place	
Bower et al., 2014	Preventive Medicine	
*Brouillette et al., 2011	Journal of Pediatrics	
Brown et al., 2007	American Journal of Public Health	
Brown et al., 2008	Journal of General Internal Medicine	
*Brown et al., 2016	American Journal of Preventive Medicine	
Brown et al., 2018	Int. Journal of Environmental Research Public Health	
Browning & Cagney, 2002	Journal of Health	
Browning & Cagney, 2003	Journal of Health and Social Behavior	
Buehler et al., 2019	Preventing Chronic Disease	
*Burns & Inglis, 2007	Health & Place	
Caughy et al., 2003	Social Science & Medicine	
Cerin et al., 2006	Medicine & Science in Sports & Exercise	
*Chen et al., 2006	Journal of Urban Health	
Chuang & Gober, 2015	Environmental Health Perspectives	
Clarke et al., 2015	Annals of Epidemiology	
Cohen et al., 2003	American Journal of Public Health	
Comstock et al., 2010	Journal of Environmental Psychology	
*Coulton et al., 2002	Neighborhood Health Indicators	
Cremonese et al., 2010	Cadernos Saúde Pública	
Cubbin et al., 2005	Perspectives on Sexual and Reproductive Health	
Cunradi et al., 2000	Annals of Epidemiology	
Curry et al., 2008	Social Science & Medicine	

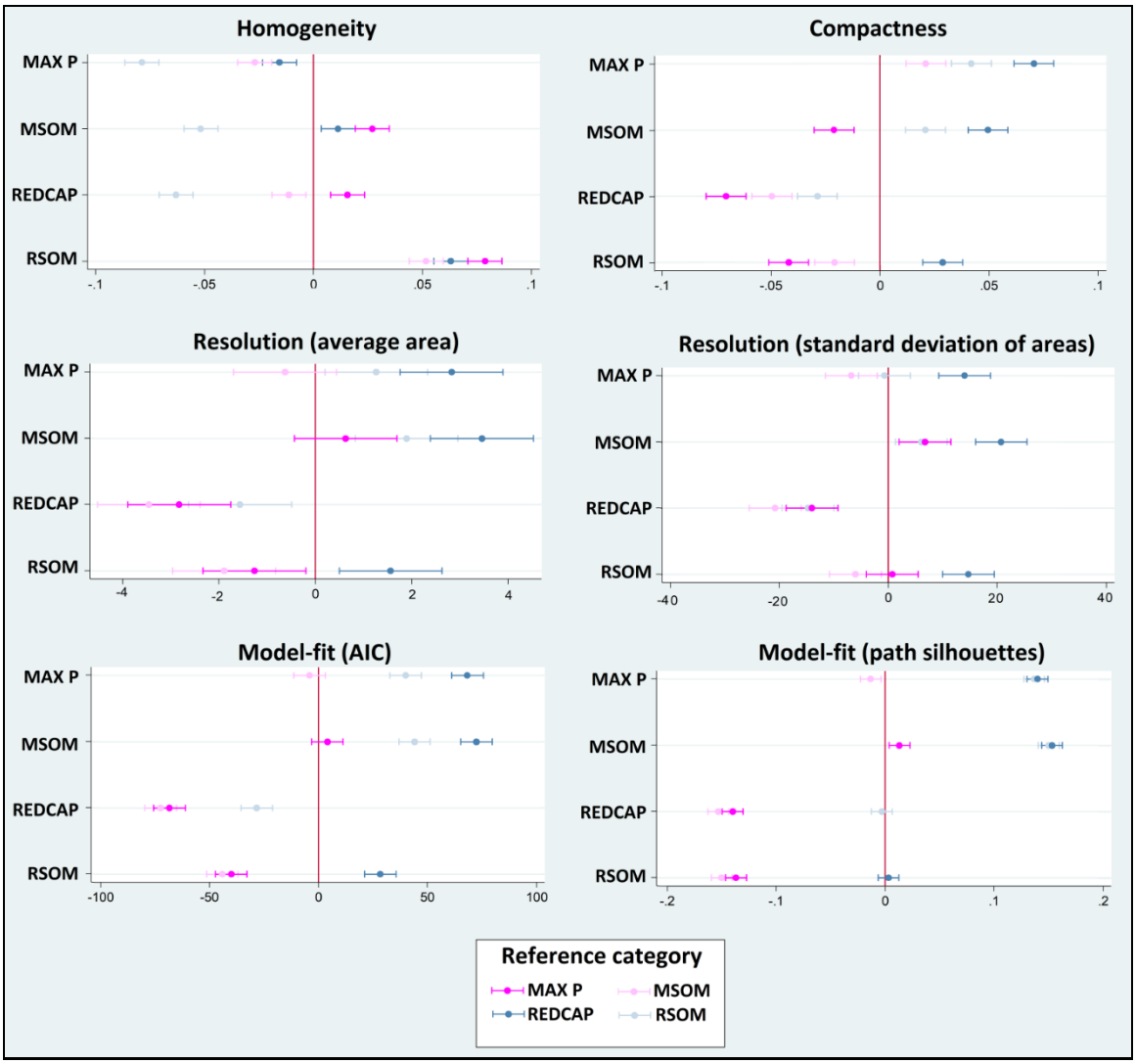
Curry et al., 2008	Social Science & Medicine
*Darden et al., 2010	Annals of the Association of American Geographers
Diez Roux et al., 2016	Global Heart
Doubeni et al., 2012	American Journal of Public Health
*Drewnowski et al., 2014	International Journal of Obesity
*Drewnowski et al., 2007	Social Science & Medicine
*Drewnowski et al., 2014	Preventing Chronic Disease
Dubowitz et al., 2012	Obesity
Duncan et al., 2014	American Journal of Epidemiology
*Duncan et al., 2013	Journal of Urban Health
Dury et al., 2016	Journal of Applied Gerontology
Echeverría et al., 2008	Health & Place
Elliott, 2000	Health & Place
*English et al., 2003	Social Science & Medicine
Eschbach et al., 2005	Cancer
*Eschbach et al., 2004	American Journal of Public Health
Estabrooks et al., 2003	Annals of Behavioral Medicine
*Ewart & Suchday, 2002	Health Psychology
Finch et al., 2010	Health & Place
*Foster et al., 2016	Journal of Urban Affairs
*Franco et al., 2008	American Journal of Preventive Medicine
Frank et al., 2006	Journal of the American Planning Association
Frank et al., 2007	Journal of Health and Social Behavior
Freeman et al., 2012	Journal of Urban Health
Freeman et al., 2011	Cancer Epidemiology Biomarkers & Prevention
*Freeman Anderson, 2020	City and Community
Friche et al., 2013	Journal of Urban Health
Gauvin et al., 2008	American Journal of Epidemiology
Giatti et al., 2010	Social Science & Medicine
Gibbons et al., 2018	PLOS ONE
*Gibbons et al., 2019	PLOS ONE
Giovenco et al., 2016	Health & Place
*Giovenco et al., 2018	Nicotine & Tobacco Research
Gomez & Muntaner, 2005	Critical Public Health
*Goodman et al., 2016	Clinical Orthopedics & Related Research
*Grant et al., 2018	International Journal of Environmental Research Public Health
Greene et al., 2015	American Journal of Public Health
*Greves Grow et al., 2010	Social Science & Medicine
Gyimah-Brempong, 2001	Southern Economic Journal
Haley et al., 2018	Sexually Transmitted Diseases
Hannon & Cuddy, 2006	The American Journal of Drug and Alcohol Abuse
*Harlan et al., 2013	Environmental Health Perspectives

Hatzenbuehler et al., 2015	American Journal of Public Health
Henriksson et al., 2010	Social Science & Medicine
*Henry et al., 2013	Health & Place
Höfelmann et al., 2015	Cadernos de Saúde Pública
*Hu et al., 2020	Journal of Urban Health
Hu et al., 2020	Preventive Medicine
*Huang et al., 2019	Int. Journal of Environmental Research Public Health
Hunt et al., 2007	Social Psychology Quarterly
Hybels et al., 2006	The American Journal of Geriatric Psychiatry
Inagami et al., 2009	Journal of Urban Health
Inagami et al., 2007	Social Science & Medicine
Inagami et al., 2006	American Journal of Preventive Medicine
Islam et al., 2006	Health Economics
James et al., 2015	Inter Journal of Environmental Research & Public Health
*Ji et al., 2020	BMC Public Health
Johns et al., 2012	Social Psychiatry and Psychiatric Epidemiology
Jokela, 2014	American Journal of Epidemiology
Jolly et al., 2010	The Journal of Rheumatology
Juhn et al., 2005	Social Science & Medicine
*Kardan et al., 2015	Scientific Reports
Keyes et al., 2012	Drug and Alcohol Dependence
Keyes et al., 2012	Drug and Alcohol Dependence
*Kim et al., 2019	Preventing Chronic Disease Public Health Research, Practice, Policy
Kim, 2010	Social Science Research
*Kind et al., 2014	Annals of Internal Medicine
Kirby & Kaneda, 2005	Journal of Health and Social Behavior
Kneeshaw-Price et al., 2015	Journal of Urban Health
Kobetz et al., 2003	Health & Place
*Koh et al., 2015	Applied Geography
*Kolak et al., 2019	Preventing Chronic Disease
Kowaleski-Jones et al., 2013	International Journal of Environmental Health Research
*Kramer et al., 2010	International Journal of Health Geographics
Kravdal, 2006	Health & Place
Kreuter et al., 2006	Health Education & Behavior
Krieger et al., 2017	Journal of Urban Health
Krieger et al., 2002	American Journal of Public Health
Kruger et al., 2007	American Journal of Community Psychology
Kubzansky et al., 2005	American Journal of Epidemiology
Lamichhane et al., 2015	American Journal of Epidemiology
Lash, 2003	American Journal of Epidemiology
Laveist & Wallace, 2012	Race, Ethnicity, and Health
Lee et al., 2008	American Sociological Review

Lee & Cubbin, 2002	American Journal of Public Health
Lee et al., 2009	American Journal of Public Health
*Lee et al., 2017	Int. Journal of Environmental Research & Public Health
Li et al., 2010	American Journal of Public Health
Lindberg & Orr, 2011	American Journal of Public Health
Litt et al., 2011	American Journal of Public Health
Lopez, 2007	Obesity
Maass et al., 2016	Health & Place
*MacQuillan et al., 2017	Journal of Community Health
Main et al., 2011	Preventing Chronic Disease
Masi et al., 2007	Social Science & Medicine
*McClure et al., 2019	Health & Place
Merlo et al., 2019	Health & Place
Messer et al., 2006	Annals of Epidemiology
*Messer et al., 2012	Health & Place
Meyers et al., 2013	Translational Psychiatry
Meyers et al., 2013	Translational Psychiatry
Mohnen SM et al., 2012	BMC Public Health
Mohnen et al., 2011	Social Science & Medicine
Molina et al., 2012	Drug and Alcohol Dependence
*Mooney et al., 2017	Practice of Epidemiology
Mooney et al., 2017	American Journal of Epidemiology
Mooney et al., 2014	Epidemiology
Moore & Diez Roux, 2006	American Journal of Public Health
*Morgenstern et al., 2009	Annals of Neurology
Morland et al., 2002	American Journal of Preventive Medicine
Mujahid et al., 2008	Epidemiology
Neckerman et al., 2009	Journal of Public Health Policy
Nelson et al., 2017	Journal of General Internal Medicine
*Nguyen et al., 2016	JMIR Public Health and Surveillance
*Nguyen et al., 2017	Scientific Reports
Nordstrom et al., 2004	The cardiovascular health study
*Patel et al., 2003	Annals of Epidemiology
Patterson & Chapman, 2004	American Journal of Health Promotion
Pearl et al., 2001	American Journal of Public Health
Pearlman et al., 2003	Public Health Reports
*Pedigo et al., 2011	PLoS ONE
*Piccolo et al., 2015	Social Science & Medicine
Pickett, 2002	Annals of Epidemiology
Powell et al., 2006	American Journal of Public Health
Powell et al., 2007	Preventive Medicine
Prentice, 2006	Social Science & Medicine
Prochaska et al., 2020	Preventive Medicine

Putrik et al., 2015	Journal of Urban Health
Putrik et al., 2019	BMC Public Health
Redelings et al., 2010	Journal of Urban Health
Reichman et al., 2009	Health & Place
Rios et al., 2012	Annals of Behavioral Medicine
Roh et al., 2011	Journal of Immigrant and Minority Health
Ross , 2000	Social Science & Medicine
Ross, 2000	Journal of Health and Social Behavior
Ross & Mirowsky, 2001	Journal of Health and Social Behavior
Roux et al., 2001	New England Journal of Medicine
*Rowe et al., 2016	Journal of Urban Health
Rudolph et al., 2014	Social Psychiatry and Psychiatric Epidemiology
Rundle et al., 2008	Social Science & Medicine
*Rundle et al., 2009	Environmental Health Perspectives
Russell et al., 2012	Cancer Causes & Control
Saelens et al., 2012	American Journal of Preventive Medicine
Sallis et al., 2009	Social Science & Medicine
*Santos et al., 2010	International Journal of Health Geographics
Sasson et al., 2010	Annals of Internal Medicine
*Sasson et al., 2011	Resuscitation
Sasson et al., 2012	New England Journal of Medicine
Schaefer-McDaniel, 2009	Health & Place
Schulz et al., 2000	Journal of Health and Social Behavior
Shacham et al., 2013	HIV Medicine
Shai, 2006	Public Health Reports
*Shi et al., 2016	Journal of Addiction
Shimotsu et al., 2013	Drug and Alcohol Dependence
Sierra et al., 2015	Journal of Investigative Dermatology
Silver et al., 2002	Social Science & Medicine
*Sohn, 2013	Journal of Korea Spatial Information Society
Spring, 2018	The Gerontologist
*Stansfield & Doherty, 2019	Social Science Research
Stenger et al., 2014	Sexually Transmitted Diseases
Stjärne et al., 2006	Epidemiology
Stockdale et al., 2007	American Journal of Preventive Medicine
Subramanian et al., 2006	Journals of Gerontology
Subramanian et al., 2005	American Journal of Public Health
Suruda et al., 2005	BMC Health Services Research
Tam et al., 2014	Influenza and Other Respiratory Viruses
*Thomas et al., 2008	Journal of Urban Health
Tolsma et al., 2009	Acta Politica
Tonne et al., 2005	Circulation
Unger et al., 2014	Circulation

*van Holm et al., 2020	Journal of Public Health
*Vanderslice & Fulton, 2012	Med Health RI
Vernez Moudon et al., 2011	American Journal of Preventive Medicine
Vinikoor-Imler et al., 2011	Social Science & Medicine
Volkova et al., 2008	Journal of the American Society of Nephrology
*Vujcic et al., 2016	The Journal Agriculture and Forestry
Vutien et al., 2019	Digestive Diseases and Sciences
Wang, 2020	Cities
*Wang et al., 2015	Health Assessment
*Wang & Immergluck, 2018	American Journal of Public Health
Warren Andersen et al. 2018	American Journal of Preventive Medicine
Weden et al., 2008	Social Science & Medicine
*Weiss et al., 2011	Journal of Urban Health
*Weiss et al., 2007	American Journal of Preventive Medicine
Wen et al., 2003	Social Science & Medicine
Wen et al., 2005	Health Services Research
*Wen & Christakis, 2006	Social Science & Medicine
White & Borrell, 2006	Ethnicity and Disease
Wight et al., 2009	The Journals of Gerontology
Wright & Kloos, 2007	Journal of Environmental Psychology
Xue et al., 2005	Archives of General Psychiatry
Yousey-Hindes & Hadler, 2011	American Journal of Public Health
*Zhang et al., 2011	Health & Place



5. 1 Pairwise comparisons of adjusted linear predictions of regionalization approach with 95% confidence intervals for mean homogeneity, compactness, region size, region variability, AIC, and path silhouette score. Note that error bars reflect one way ANOVA with a Bonferroni correction.

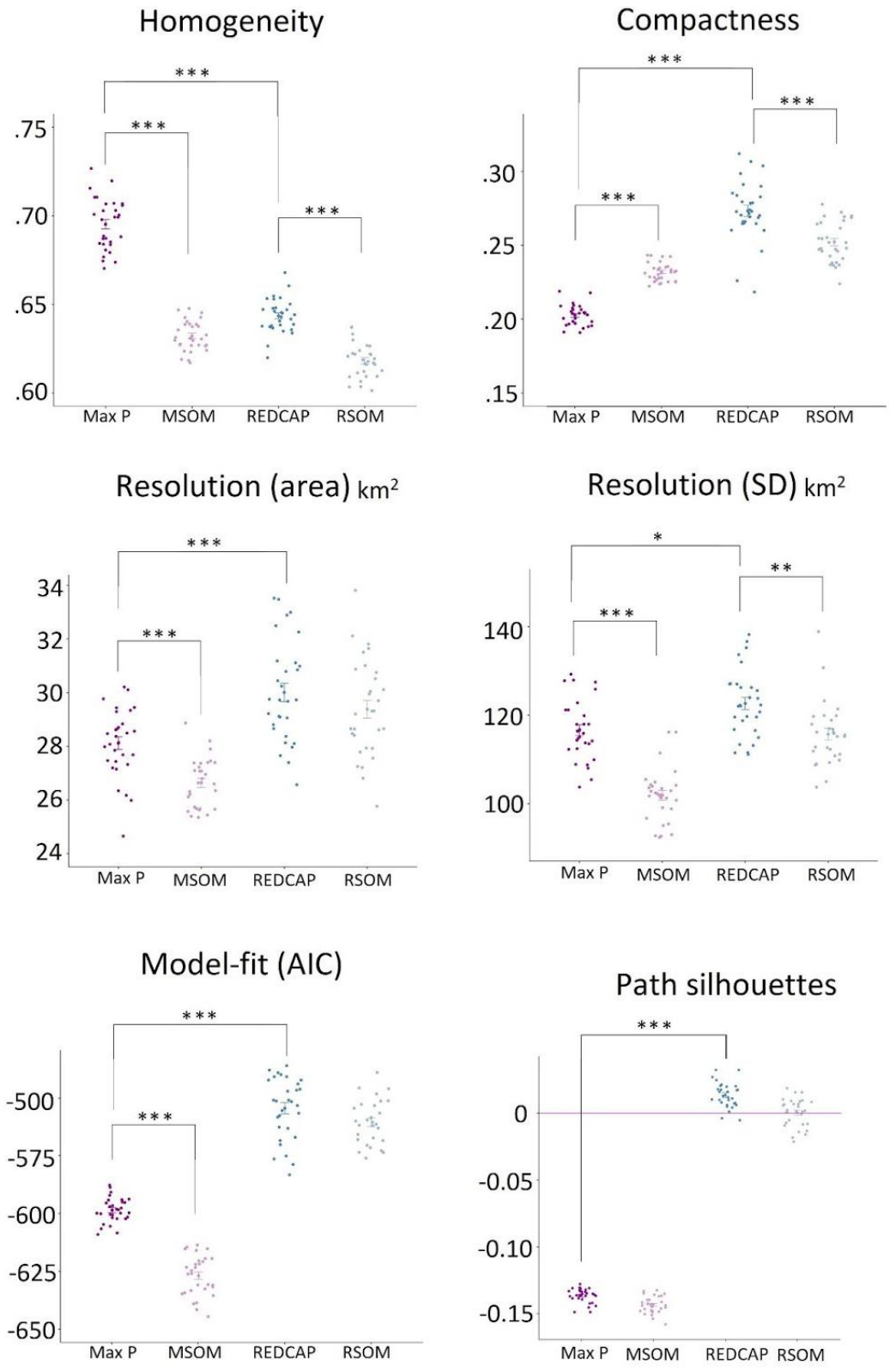


Figure 5.2. Scatter plots and standard error bars for the 30 runs of Max P, MSOM, REDCAP, and RSOM for the 6 different assessment measures. Significance is taken from Games-Howell post hoc test or, for path silhouettes, Dunn's test. Alpha levels were adjusted according to the Bonferroni correction.